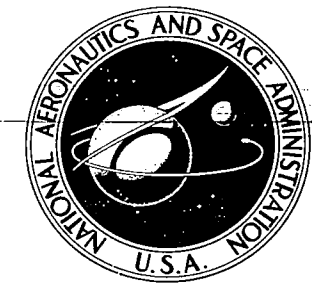
NASA CONTRACTOR REPORT



APPLICATION OF STATISTICAL ASSOCIATION TECHNIQUES FOR THE NASA DOCUMENT COLLECTION

by Paul E. Jones, Robert M. Curtice,
Vincent E. Giuliano, and Murray E. Sherry

Prepared by
ARTHUR D. LITTLE, INC.
Cambridge, Mass.
for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • FEBRUARY 1968

APPLICATION OF STATISTICAL ASSOCIATION TECHNIQUES FOR THE NASA DOCUMENT COLLECTION

By Paul E. Jones, Robert M. Curtice, Vincent E. Giuliano, and Murray E. Sherry

Prepared under Contract No. NASw-1051 by
ARTHUR D. LITTLE, INC.
Cambridge, Mass.

for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – CFSTI price $3.00

ABSTRACT

Computer programs for batched associative search of the NASA coordinate-indexed document collection have been prepared and tested. Cooccurrence counts for the approximately 18,000 terms used to index about 100,000 documents were obtained and association matrices based on various subsets of these counts were developed. One Section of the report describes the system as a whole; the remainder is focused on a discussion of the behavior of the system observed during tests of the programs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES
(continued)

PART I

APPLICATION OF STATISTICAL ASSOCIATION TECHNIQUES FOR THE NASA DOCUMENT COLLECTION

SUMMARY

A. Background

When our work on this project began, the possibility of using statistical association techniques in document searching operations and other mechanized documentation activities was receiving considerable public attention. Numerous research groups were active in the area. Many of these were represented at the Conference on Statistical Association Techniques for Mechanized Documentation[1] which was held in Washington, D.C., at about that time. Several hundred investigators were present to discuss the results of efforts over the preceding few years.

Since those early days of imaginative exploratory investigation, the pace of publication and the rate of apparent progress has noticeably slackened. Indeed, it has seemed to some that investigation of the subject has largely been completed, even abandoned. But the external appearances are misleading. The activity in the field has recently been devoted to consolidation -- in our case, a gearing up to treat the problems of very large document collections. This developmental work, which is of little interest in itself, is nevertheless an inevitable step in moving research ideas to the point of practical test.

The principal objective of the work reported here was to create a set of computer programs capable of practical associative processing of NASA's growing document collection. The secondary objective was to process the collection with these programs, exhibiting the achievement of a capability to perform associative search of the collection. Finally, a third objective was to learn what we could about the use of statistical associations in a large collection, by assessing the experience gained during limited test operation of the system.

Our work, though delayed, has been successful on all three counts. A system of programs for associative processing in large collections is in existence. An exhaustive file of term co-occurrence counts (all terms used in nearly 100,000 documents) has been prepared. Associative search of the collection has been conducted. The system stands ready for test, with extensive facilities built-in for flexibly adjusting the parameters of the system (association formula, intervention strategy, mode of use, etc.) to accommodate the needs either of investigators or of system implementers.

This report is organized in three parts. The Summary (Part I) itemizes principal resources that have been created and states our conclusions. Part II illustrates the operation of the associative retrieval system and related topics, and Part III contains the detailed technical description of the complete system we prepared.

B. Tasks

1. Program All Associative Processing System for NASA. - The complete system contains about 20 major programs and a dozen minor ones. The IBM 7090/94 is used for centralized file maintenance and the IBM 1401/1410 is used for decentralized searching operations. This system can be applied to most existing large-scale co-ordinate-indexed retrieval collections to provide a supplementary associative processing capability. The programs are directly applicable to systems which are organized for tape searching of a Linear File.

The four principal components of this system of programs are summarized in the attached figure and below:

i) File Generation (Conversion)

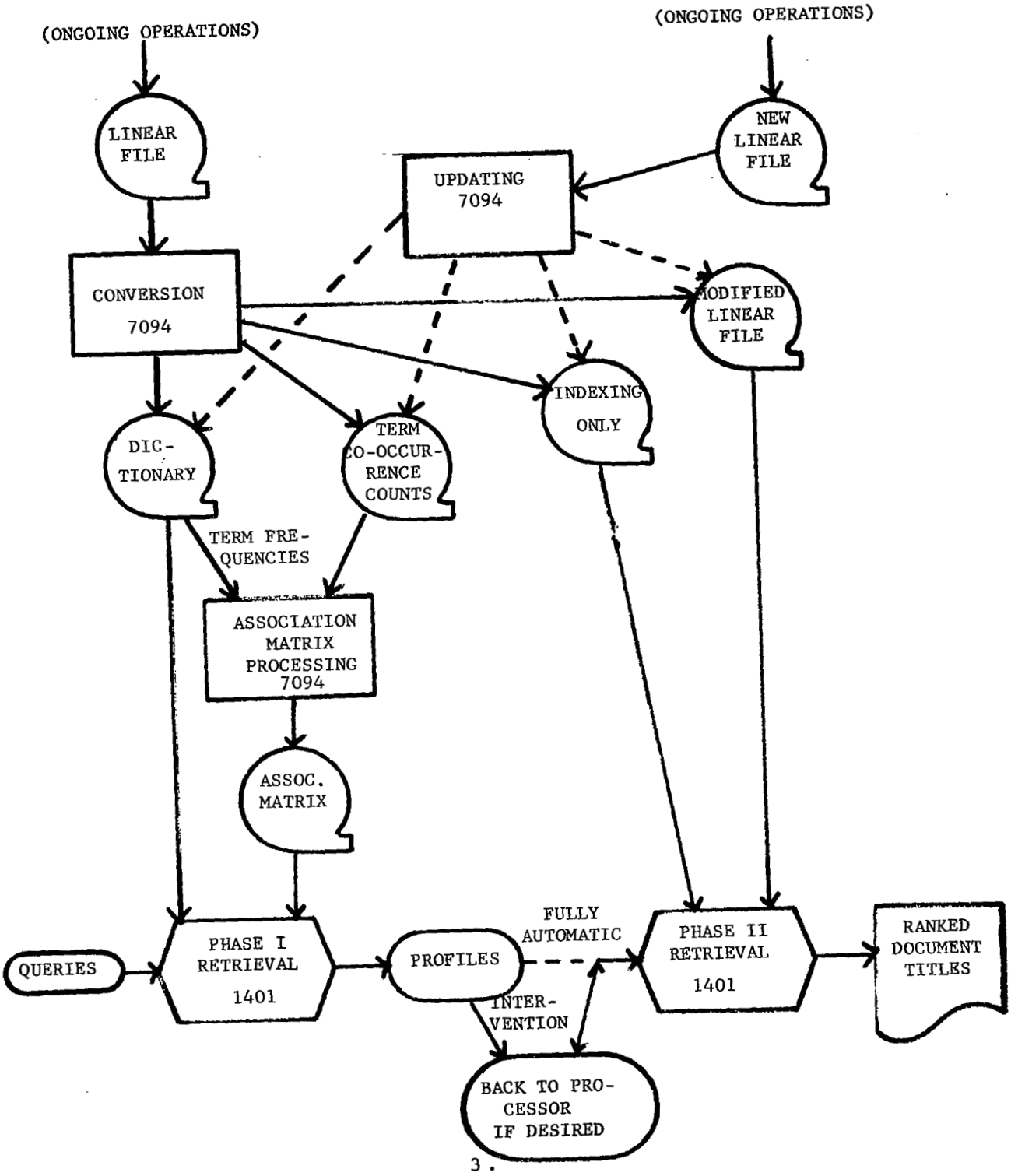
The programs accept the Linear File -- a set of sequentially-ordered term-indexed document citations on magnetic tape -- and produce four principal outputs: (a) a re-formatted Linear File which contains the document-descriptive material only (not the terms), (b) a compact encoding of the assignments of index terms (employing term numbers), (c) dictionaries of the terms used, recording their usage frequencies, and (d) a file of the term co-occurrence counts for all pairs of terms. These programs will handle collections of one million documents indexed by up to 32,000 different terms.

ii) Updating Programs

To accommodate new accessions to the collection, a related set of programs is provided which converts the new segment of the Linear File to the formats used in this system and updates the summary files: the term-usage counts are augmented, the dictionaries used are supplemented by new terms introduced, and the term co-occurrence counts are updated. These programs for file maintenance are intended for centralized operation and principally use the IBM 7090/94.

iii) Association Matrix Processing

The term co-occurrence counts, together with term usage frequency data are meant to be periodically subjected to Association Matrix processing. For this purpose, programs have been prepared for flexibly selecting portions of the complete file of co-occurrence data for processing. The capability to calculate both first and higher generation associations has been included. Because the size of the data base is very large, these programs have had to be designed both for investigative flexibility and good operating efficiency. Association matrices of dimension 3000 x 3000 can be processed by all programs. In addition, most of the programs are available in a revised version in which this limitation is removed to permit the use of matrices of dimension up to 32,000. Both rectangular


(ONGOING OPERATIONS)
(ONGOING OPERATIONS)
LINEAR FILE
NEW LINEAR FILE
UPDATING 7094
CONVERSION 7094
MODIFIED LINEAR FILE
INDEXING ONLY
DIC-TIONARY
TERM CO-OCCUR-RENCE COUNTS
TERM FRE-QUENCIES
ASSOCIATION MATRIX PROCESSING 7094
ASSOC. MATRIX
QUERIES
PHASE I RETRIEVAL 1401
PROFILES
FULLY AUTOMATIC
INTER-VENTION
PHASE II RETRIEVAL 1401
RANKED DOCUMENT TITLES
BACK TO PRO-CESSOR IF DESIRED

3.

and square matrices can be employed and a variety of association formulas can be incorporated in lieu of those now programmed. These programs all operate on the IBM 7094.

iv) Retrieval Programs

Designed for decentralized operation on the IBM 1401, the retrieval process operates in two phases. The first phase computes association profiles in response to queries that are presented as (unstructured) sets of weighted terms. Multiple requests are processed concurrently, and the output profiles are presented as listings of associated terms, arranged with the most closely associated terms at the head of the list. This program uses the association matrix centrally developed on the IBM 7090/94 and term dictionaries that are centrally kept up-to-date. Association matrices of dimension up to 32,000 can be used in generating association profiles with the 1401 retrieval program; matrices beyond 6,000 rows are cumbersome on the 1401.

An option for intervention is provided between Phase I and Phase II. At this point, the profile can be inspected and the most desirable associations selected before proceeding with the document search phase. Phase II, also conducted on the IBM 1401, uses the profiles prepared in Phase I for weighting documents, and the final product is a ranked listing of selected document-defining material drawn from the original Linear File. This program is designed for efficiency in using the small, slow, IBM 1401 machine for searching a very large collection. Programmed limits are one million documents indexed by 32,000 terms. While searches of 100,000 documents are time-consuming with the 1401 computer, they are feasible. The existing programs also operate on faster machines, however.

2. Process the NASA Document Data Base. - The entire NASA collection in existence at the initiation of our work (almost 100,000 documents) was processed by the conversion program. The principal resulting files were:

a) Term co-occurrence counts for all the terms (about 18,000) in the NASA vocabulary, including the "published" terms. This is a major file and is the most important body of data generated during this work.

b) Dictionary tapes (and listings) showing term usage frequencies for all terms, both alphabetically-ordered and frequency-ordered.

c) A square Association Matrix (1000 x 1000) was derived from the co-occurrences of the most frequent terms in the collection. This matrix was subsequently used for retrieval tests on the 1401. The Association formula used was $fab/fa \cdot fb$ where fab is the co-occurrence count of two terms a and b, and fa, fb are the individual term frequencies. Over a hundred printed profiles for short, medium, and long requests were derived from this association matrix. Samples of output appear in Part II.

d) A rectangular Association Matrix (100 x 18,000) showing full profiles for the terms in the frequency interval 222 - 244. The same association formula as before was used. Several dozen printed profiles were derived from this matrix and studied. Two profiles in particular were studied in depth and rearranged into alphabetic, frequency-ordered, co-occurrence count ordered, and other listings for inspection. Samples of output appear in Part II.

e) Another rectangular Association Matrix (88 x 18,000) was prepared for 88 terms individually selected by NASA as interesting for thesaurus-study purposes. The Association formula used was $fab/\sqrt{fa \cdot fb}$; terms with low co-occurrence counts and low frequency were eliminated. A special printout of the 88 x 18,000 co-occurrence count matrix was prepared for reference. A special profile listing of the top 200 associated terms for each of the 88 terms was also prepared. Samples of output are given in Part II.

f) An Associative Search of the 100,000-document collection was conducted using a live request that had been processed routinely by the NASA Facility. Part II contains a description.

3. Limited Test Operation. - Trial operations were a small but significant portion of our efforts. They were aimed principally at determining that the programs were producing the behavior expected of an associative processing system. The test operations showed the following:

i) They demonstrated the capability of the associative process to expand the vocabulary of a search prescription, to handle unedited full-text requests from the original requestor, and to suggest terminology suited to narrowing a search and making it specific.

ii) They demonstrated the feasibility of conducting associative search in a very large collection, highlighting the great ease with which term profiles can be prepared.

iii) They helped to clarify the adjustments and choices of parameters applicable to using the techniques within NASA's operational context.

The test operations also involved several subsidiary investigations of interest which included:

iv) A preliminary investigation of the effect of different association formulas on the character of profiles generated for the very large (and occasionally very technical) NASA vocabulary.

v) A preliminary investigation of the usefulness of second-generation associations within the NASA context.

Discussions of all of these topics appear in Part II.

C. Resources Now Available

Our work has resulted in the creation of significant resources whose exploitation in NASA's operational activities, as well as within operational and developmental documentation activities in other organizations, deserves situational appraisal:

1. The availability of a comprehensive and well-tested set of computer program packages for associative retrieval processing of any available coordinate-indexed collection is noteworthy. The total programming and developmental investment in these packages represents an investment which need not be duplicated by other organizations wishing to experiment with associative processing.

2. The well-developed set of IBM 1401-compatible associative searching programs now in use is the third re-design of this portion of the program package. It is well suited to field testing of associative processing using the NASA collection, and to operation by interested parties without extensive training.

3. The set of term association measures computed (or readily computable) among the 18,000 terms in the NASA collection is a major resource. The data set is potentially useable within such undertakings as (a) automatic thesaurus compilation, (b) association-guided searching, (c) studies of indexing vocabularies, and (d) work on selective dissemination systems.

4. Sets of supporting computer programs and related studies, each of which has potential relevance to future applications. A number of initial investigations have been made into related areas of association-aided document retrieval, and these may help provide a base on which to build more careful explorations. For example, programs which automatically index abstracts and feed the resulting "Linear File" into the present associative system have been prepared and the results evaluated. (See Ref. 12)

D. Findings

1. Overall. - Our observations during trials of the system's operation have reinforced our view that the use of term associations holds great promise for improving the cost/effectiveness of retrieval searching. While our operating experience is limited, the application of completely automatic associative techniques in a very large mechanized retrieval system is now an accomplished fact. Evaluation of the success of this methodology in providing solutions to diverse problems that arise in operational document-searching contexts has never been done. Executing such an evaluation is a necessary next step in the near future, for until it is done, research groups in the field are forced to restrict themselves to estimates and judgments that are easily dismissed as the pronouncements of advocates of the given approach. Without the collaboration of an operational documentation service organization there is little prospect of progressing beyond the stage now achieved, where professional investigators exhibit -- with difficulty -- what they judge to be useful attributes of the procedure. Our findings are our own judgments; we have no special knowledge of the operational needs or problems of the operators of document-searching systems that use the NASA collection.

2. Key Judgments

i) Use of statistical associations for machine-aided thesaurus compilation appears exceptionally promising. The unprecedented richness of the associations presented in machine-prepared profiles (involving both synonyms and other related words) deserves to be exploited further.

ii) The best way to use term associations is in conjunction with, not as replacement for, human query renegotiation.

iii) The ready availability of rich association lists reduces the need for expertise in query formulation. Our observations consistently support the use of statistical associations in a time-sharing query environment, where information-system users are urged to conduct their own interactive searching.

iv) Certain technical expectations, such as the possible utility of second versus first-order associations, have not yet been resolved and bear further study. We do not, however, see these issues as affecting the perceived operational utility of the associative tools.

3. Practicality

i) The data processing involved in computing co-occurrence counts for a collection of 100,000 documents is arduous with the IBM 7094 system. But once done it need not be repeated.

ii) Association Matrix Processing with the limits of the existing program package (e.g., up to 3000 x 3000 square matrices) is easy. It is simply a matter of submitting computer runs.

iii) The Phase I retrieval operation (which produces term profiles on the 1401) is relatively fast (one hour using a 1000 x 1000 matrix), very easy and also inexpensive if multiple queries are batched for simultaneous running.

iv) The Phase II (Document Retrieval Phase) is time-consuming on the 1401 when 100,000 documents are searched (six hours) but easy, convenient, straightforward and not expensive when only 10 - 20,000 documents are searched at a time.

PART II

OBSERVATIONS FROM TRIAL OPERATIONS

A. Introduction

Research investigations into the use of statistical features of term co-occurrence in a mechanized document searching system have been conducted by numerous groups of individuals since 1958. The influence among others, of Stiles (3, 4), Maron, Kuhns and Ray (5), and Doyle (6) in stimulating this line of investigation is widely acknowledged.*

Up to the initiation of the present work, the idea of applying statistical association measures in mechanized documentation systematically unfolded along the usual progression that carries an idea into practice. Stages of research, clarification, early evaluation, and laboratory testing on miniature and medium scale prototype systems can be identified in the basic work that has gone on. The present work constitutes the next step in such a progression and represents the scaling up of well-studied prototypes for application in a real-life situation.

In the broadest view, our job was to move a set of research ideas out of the laboratory -- to solve all the practical problems involved in doing so, to exercise our best judgment as to the proper way to proceed, to design, test, and, in short, bring into existence an associative searching system applicable to the NASA collection. We stopped just short of using the system in practice; rather, we endowed NASA, and of course other agencies, with the capability to try associative searching for their purposes, placing them in a position to assess the usefulness of a concrete system in practice.

We have outlined the resulting system in the Summary. For those readers who are interested in the programs and their functioning, a complete description of the system is given in Part III of this report. Our objective in this part of the report, however, is to demonstrate some of the uses of associations and to present some of the most interesting observations that have emerged from trials using them.

*In particular, our own early work 7 - 12 was significantly influenced by these others. Further background and recent developments are provided in (15 - 30).

PART II

OBSERVATIONS FROM TRIAL OPERATIONS

A. Introduction

Research investigations into the use of statistical features of term co-occurrence in a mechanized document searching system have been conducted by numerous groups of individuals since 1958. The influence among others, of Stiles (3, 4), Maron, Kuhns and Ray (5), and Doyle (6) in stimulating this line of investigation is widely acknowledged.*

Up to the initiation of the present work, the idea of applying statistical association measures in mechanized documentation systematically unfolded along the usual progression that carries an idea into practice. Stages of research, clarification, early evaluation, and laboratory testing on miniature and medium scale prototype systems can be identified in the basic work that has gone on. The present work constitutes the next step in such a progression and represents the scaling up of well-studied prototypes for application in a real-life situation.

In the broadest view, our job was to move a set of research ideas out of the laboratory -- to solve all the practical problems involved in doing so, to exercise our best judgment as to the proper way to proceed, to design, test, and, in short, bring into existence an associative searching system applicable to the NASA collection. We stopped just short of <u>using</u> the system in practice; rather, we endowed NASA, and of course other agencies, with the capability to <u>try</u> associative searching for their purposes, placing them in a position to assess the usefulness of a concrete system in practice.

We have outlined the resulting system in the Summary. For those readers who are interested in the programs and their functioning, a complete description of the system is given in Part III of this report. Our objective in this part of the report, however, is to demonstrate some of the uses of associations and to present some of the most interesting observations that have emerged from trials using them.

---

*In particular, our own early work 7 - 12 was significantly influenced by these others. Further background and recent developments are provided in (15 - 30).

Though the development of these observations has been a small portion of our actual effort, and though we wish a less preliminary picture could be presented, we feel that an informal description of how the system works is of general interest. Accordingly, we devote this part of the report to treating three topics which arise in the use of statistical associations in a retrieval context, and we use these discussions as a vehicle for demonstrating the use of the system in practice.

In Section B we comment on the size of the co-occurrence count data base we used and on the practicalities of dealing with it. Our choices of submatrices (as well as of formulas and other parameters) govern all the observations recorded here.

Section C is devoted to carrying a full text (original) request through Phase I retrieval and shows the term association profile developed for an unedited request. Section D reports, on the other hand, a fully automatic associative search through the 100,000 document collection using an edited request prepared by a search analyst within NASA's Scientific and Technical Information Facility. In both these sections, the association matrix is of dimension 1000 x 1000.

In sections E and F the discussion becomes more technical. Attention turns towards profiles developed over not 1000 high-frequency terms but over all (approximately) 18,000 terms in the full vocabulary. Section E addresses the very difficult problem of studying second-generation associations. Section F treats some of the problems of selecting an association formula for thesaurus-building purposes.

Throughout, the illustrative examples are drawn from outputs obtained during our test operation of the system.

B. Co-occurrence Count Data

The association process critically depends on the centrally-gathered term co-occurrence-count data. While the user of the associative retrieval system is indifferent to the centralized operations that go into constructing the necessary files, we describe this file briefly to serve as a starting point.

The principal body of data we developed consisted of all the co-occurrence counts for all the terms used to index the approximately 100,000 documents that were in the NASA collection when we started.

An array of the co-occurrence counts is conveniently organized in the form of a matrix where the rows and the columns correspond to the terms in the vocabulary, and the number of documents in which two terms co-occur is recorded at the intersection of the row and column. Since

the whole NASA vocabulary contained over 18,000 terms*, exhaustive co-occurrence counting leads to an 18,000 x 18,000** matrix of co-occurrence counts.

This matrix -- just the numbers, recorded compactly -- requires more than two full reels of magnetic tape for storage of the non-zero elements. It takes about 20 minutes for an IBM 7094 computer merely to read through a file of this size. Because processing costs are high, attention has to be focused on using pieces of this file. The existence of the exhaustive file is very significant in permitting complete flexibility in choosing which piece to study and use. But some choice is a practical necessity (14).

In pursuing our initial investigations, we have made use of a package of programs (Part III) which work upon the 18,000 x 18,000 co-occurrence count matrix. Those programs permit the flexible creation of a variety of association matrices as needed and provide the basic capability to isolate a segment of his huge matrix for intensive processing. This processing involves not only the computation of association coefficients, but also the deletion of matrix elements that are too small, the adjustment of diagonal elements, multiplying (certain) matrices, making necessary format adjustments, and editing the matrices for display of association profiles. The basic investigative capability to try out a variety of ideas on a smaller scale is available. We have tried only a few of the interesting possibilities. Our selections will be clear from the discussions which follow.

---

*Throughout this report we treat the collection as having 100,000 documents (though the exact number is 95,509) and the vocabulary as having 18,000 terms (though the exact number is 18,292). These numbers change with collection growth and have no particular significance. For example, the programs are not limited until the numbers are far larger (one million documents, 32,000 terms). For the record, we comment that although we have come to think of the co-occurrence-count matrix as 18,000 x 18,000, there are actually 1654 terms from the vocabulary we were given that did not occur in the document sample we were given. Since these terms have zero co-occurrences, the interesting co-occurrence counts appear in a 16,638 x 16,638 matrix.

**Recent work by Fossum, et. al. (13) in processing 38,402 DDC documents has involved the determination and study of co-occurrence counts among the 600 most frequent terms. A comparison of the co-occurrence parameters between the two collections could be informative.

C. Use of Association Profiles for Query Expansion

Operation of the associative retrieval process in its most interesting application begins with a novice requestor.* Typically, he has access to a librarian who is in a position to give him limited assistance on-site in discovering at least a few terms that are vocabulary entries of the NASA system. Published lists of such terms are available and are readily consulted to get a few descriptive terms that appear to apply. But often he does not receive the benefit of such aid, and the role of associations in helping him home in on useful terminology is of great interest. In this example we show a full text request actually received by NASA and we exhibit deficiencies in its vocabulary which were remedied by the analyst who prepared a search prescription using NASA terms. We show that the association profile** also suggested the alternate terminology and we explore this key use of the association profiles in expanding a request.

1. Query Formulation. - The request exhibited in figure 1 was actually received by NASA. Since we are going to devote the next few pages to discussing that request, a careful reading of the request will be useful. The text of this request is an explicit and detailed statement of the requestor's problem. This statement could easily have been a transcription of the requestor's own language, and there is little evidence that the librarian who wrote the letter made any effort to recast the request employing NASA terms.

2. Comparison of the Analyst's Search and the Original Request. - This request, when processed by the search analyst*** at the NASA Facility, led ultimately to the search prescription shown in figure 2.

---

*We have treated in (9) some of the problems encountered by requestors who are not experts at posing search requests to a co-ordinate-indexed collection.

**Created by processing the full text request with the 1401 retrieval program using a 1000 x 1000 association matrix over the highest frequency terms.

***The analyst's job is to understand the requestor's intent, to formulate a search using NASA terms, to screen the retrieved documents, and to insure that the references sent to the requestor are acceptably responsive. The analyst has stated that his pre-search analysis took one hour for this request, that he judged his search strategy to be "moderately loose. Some irrelevant material." His search retrieved 121 documents of which he felt 99 were worth sending to the requestor (i.e., "accepted"), for an "acceptance ratio" (the analog of precision ratio using the analyst's judgment) of 99/121 = 82%. He spent ½ hour post-editing the search to provide this service.

9 June 1966

NASA Scientific and Technical Information Facility
Reference Department
Machine Search Branch
Post Office Box 33
College Park, Maryland 20740

Gentlemen:

We are interested in obtaining information on the following subject for engineering personnel of the ... Laboratory and, therefore, would like to have a NASA literature search performed:

> Zero gravity and/or partial-gravity simulation systems. The system must provide three or more degrees of freedom, but may utilize any support method or activation procedure including counterweights, cables, hydraulic motors, or servomotors. The use of gimbal ring support frames is of particular interest. The simulation system should be capable of supporting a 300-pound load composed of a 200-pound man-space suit plus a 100-pound back-pack maneuvering unit.

Correspondence requesting clarifications for this search or otherwise regarding this request should refer to the literature search identification code noted above. We shall appreciate your assistance in servicing the information needs of our laboratory personnel.

(Complimentary closing)

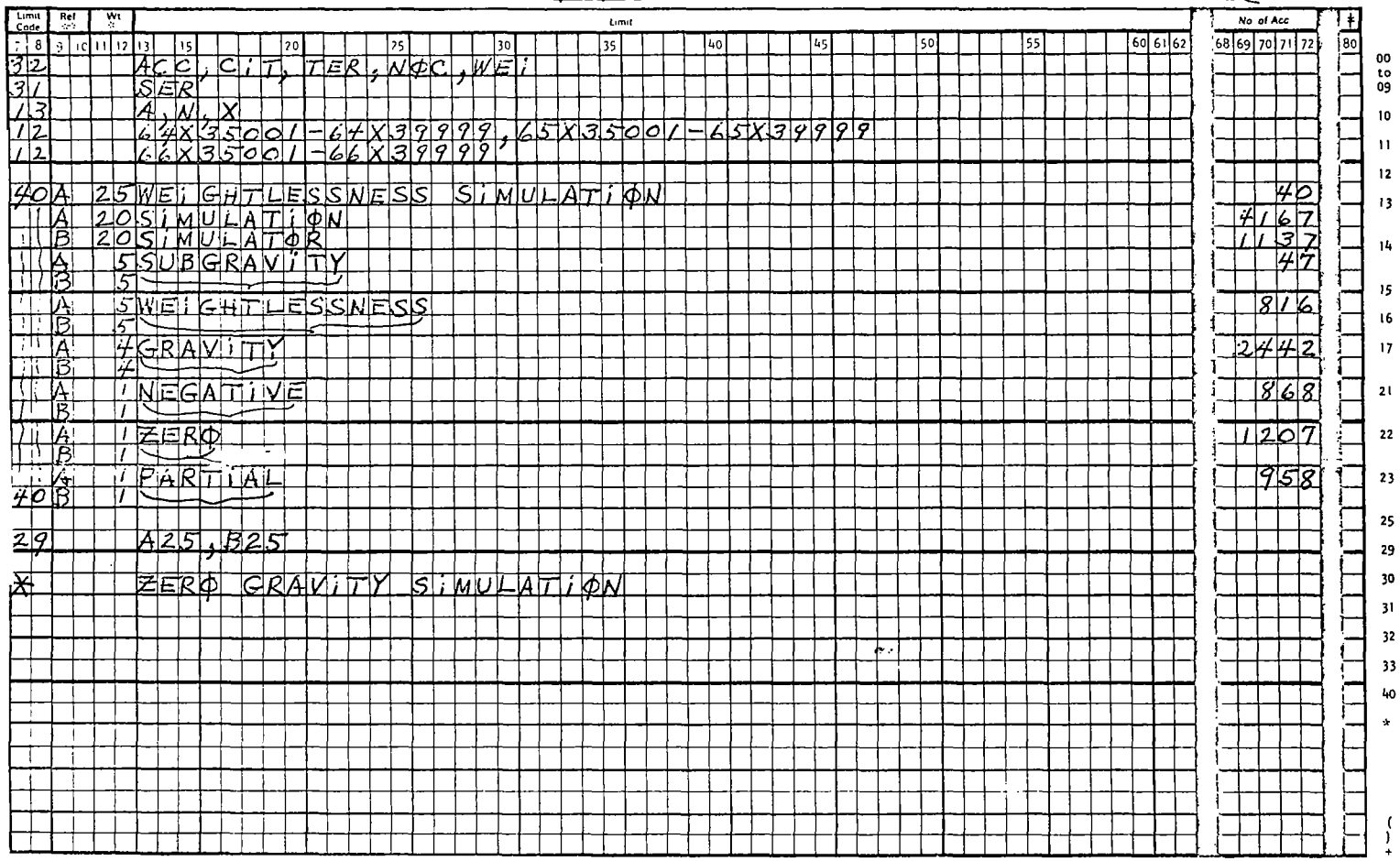
FIGURE 1 SAMPLE FULL-TEXT REQUEST

NASA LINEAR FILE SEARCH WORKSHEET
1410 Model 11 System

Problem No. 26 Bibliography No. 9000

TITLE: _____ ANALYST: Carlson Date _____

Due Date _____ No. of Terms 9 (15)



LEGEND
CODE LIMIT

00 to 09 Logical Equation
10 Security (1-6)
11 Acc. Range (Pos)
12 Acc. Range (Neg)
13 Acc. Series (A,N,X)
14 Document Type (00-09)
15 COSATI Category
16 Subject Category
17 Corporate Source (Term No.)
21 Contract No. (Root)
22 Personal Author (Root)
23 Report No. (Root)
25 No Foreign Lang.
29 Group Weights
30 Weight
31 Sort Option
32 Output Option
33 Hit Limit
40 Terms
* Comment Card

KEYPUNCH INSTRUCTIONS

(= ---0-4-8 Punch
) = 9--12-4-8 Punch
+ = 6--12 Punch

FIGURE 2 ANALYST'S SEARCH PRESCRIPTION FOR THE FULL TEXT REQUEST IN FIGURE 1

For those unfamiliar with the codes the NASA system employs for search purposes, the search the analyst chose to perform through the NASA collection on the requestor's behalf is:

(I want to see all documents indexed by:)

WEIGHTLESSNESS SIMULATION

or

SIMULATION and SUBGRAVITY
or SIMULATION and WEIGHTLESSNESS
or SIMULATION and GRAVITY and NEGATIVE
or SIMULATION and GRAVITY and ZERO
or SIMULATION and GRAVITY and PARTIAL

or

SIMULATOR and SUBGRAVITY
or SIMULATOR and WEIGHTLESSNESS
or SIMULATOR and GRAVITY and NEGATIVE
or SIMULATOR and GRAVITY and ZERO
or SIMULATOR and GRAVITY and PARTIAL

The analyst has used only the nine terms:

WEIGHTLESSNESS SIMULATION
SIMULATION
SIMULATOR
SUBGRAVITY
WEIGHTLESSNESS
GRAVITY
NEGATIVE
ZERO
PARTIAL

If we compare the analyst's terminology with that of the requestor, we notice that the analyst has omitted much of the detail which the requestor supplied. For example, the only term, among those in the analyst's Search Prescription, which conveys any notion of "devices" is the term "SIMULATOR." The requestor's catalog of "counterweights," "gimbal ring support frames" and the like (which the requestor apparently considered useful in expressing his interest) was apparently not useful for search purposes from the analyst's point of view. All this "device" terminology failed to appear in the analyst's search prescription, even though many of these words and word strings in the full text request are in fact bona fide NASA terms:

support
cable
hydraulic
motor
servomotor
gimbal
ring
frame
space suit
back pack

The difference in ways of expressing what is wanted is striking in this example. Without claiming that the example is typical of the search requests NASA receives, we do point to this example as a clear-cut case of a disparity between the way an outsider found it natural to describe his information needs and the way an "insider" found it best to express a search. If requestors are to conduct their own machine searches, it is certainly desirable and possibly crucial to develop mechanisms which provide meaningful assistance in "translating" from one natural form of expression to the other. We believe that the statistical association methodology has shown outstanding capabilities for assisting the requestor in this "translation," and we use the present example to demonstrate the point.

3. Input to the Phase I (Association Profiles) Program. - To illustrate the use of the association process for processing the requestor's statement, the first step is to obtain an association profile. The full text of the requestor's own search specification was entered into the Phase I association program. Figure 3 shows the actual input to the computer*. The requestor could have handed his manuscript to a keypuncher who knows the applicable formats and the input cards would have been prepared without further ado. Or, if he were using a remote access console, he could have transcribed the text directly.

The input "terms," that is, the individual words in the full text, in figure 3, are listed one below the other at the left. Each was assigned the same weight (viz., .9999) on input. (The retrieval program re-scales these weights to keep the numbers within manageable bounds,

---

*
Notice that it is convenient to let the machine's lookup process determine the forms of words in the term vocabulary. In this example, the keypuncher was told to enter singular and plural forms of plural nouns to avoid not finding a term for uninteresting morphological reasons.

NO. 2690 FULL TEXT-FREE

THRESHOLD .000001

TERM	INPUT WEIGHT	NORMALIZED
ZERO	9999	1125
GRAVITY	9999	1125
AND	9999	1125
OR	9999	1125
PARTIAL	9999	1125
GRAVITY	9999	1125
SIMULATION	9999	1125
SYSTEM	9999	1125
SYSTEMS	9999	1125
THE	9999	1125
SYSTEM	9999	1125
MUST	9999	1125
PROVIDE	9999	1125
THREE	9999	1125
OR	9999	1125
MORE	9999	1125
DEGREE	9999	1125
DEGREES	9999	1125
OF	9999	1125
FREEDOM	9999	1125
BUT	9999	1125
MAY	9999	1125
UTILIZE	9999	1125
ANY	9999	1125
SUPPORT	9999	1125
METHOD	9999	1125
OR	9999	1125
ACTIVATION	9999	1125
PROCEDURE	9999	1125
INCLUDING	9999	1125
COUNTERWEIGHT	9999	1125
COUNTERWEIGHTS	9999	1125
CABLE	9999	1125
CABLES	9999	1125
HYDRAULIC	9999	1125
MOTOR	9999	1125
MOTORS)	9999	1125
OR	9999	1125
SERVOMOTOR	9999	1125
SERVOMOTORS	9999	1125
THE	9999	1125
USE	9999	1125
OF	9999	1125
GIMBAL	9999	1125
RING	9999	1125
SUPPORT	9999	1125
FRAME	9999	1125
FRAMES	9999	1125

FIGURE 3 FULL-TEXT INPUT TO THE PHASE I ASSOCIATION PROGRAM SHOWING THE CASUAL EDITING PERFORMED

CONTINUED.

TERM	INPUT WEIGHT	NORMALIZED
IS	9999	1125
OF	9999	1125
PARTICULAR	9999	1125
INTEREST	9999	1125
THE	9999	1125
SIMULATION	9999	1125
SYSTEM	9999	1125
SHOULD	9999	1125
BE	9999	1125
CAPABLE	9999	1125
OF	9999	1125
SUPPORTING	9999	1125
A	9999	1125
300	9999	1125
POUND	9999	1125
LOAD	9999	1125
COMPOSED	9999	1125
OF	9999	1125
A	9999	1125
200	9999	1125
POUND	9999	1125
MAN	9999	1125
SPACE	9999	1125
SUIT	9999	1125
PLUS	9999	1125
A	9999	1125
POUND	9999	1125
BACK	9999	1125
PACK	9999	1125
MANEUVERING	9999	1125
UNIT	9999	1125

FIGURE 3 (CONTINUED)

and the normalized weights resulting from this calculation are printed for information.)

4. Output of the Phase I (Association Profiles) Program. - Figure 4 shows the association profile produced as output. (An intermediate listing -- of no interest here -- is also produced, recording the fact that "and," "but," etc.... were not in the NASA vocabulary of terms.) Since this is the first profile shown in this report, we shall dwell briefly on its format. "Batch 6" identifies this profile for reference and connects it with the input query previously shown. It also differentiates it from other batches (1, 2, 3 ... 50) that may be processed at the same time. Pages are numbered consecutively within each profile that is printed. We show only the first two pages of 15. How many terms are printed is controlled by the threshold setting (figure 3) on the input: only the terms with weight in excess of the threshold are printed. Terms are listed in decreasing order of their weight. Only terms among the 1077 most frequent terms in the NASA vocabulary were in the association matrix used in this example. The profile is thus "over" these 1077 high frequency terms. The terms from the input full text query that were in this subset of the vocabulary appear in the profile with an asterisk at the right margin. All other words in the query were completely ignored.

The term numbers shown on the profile are those we have assigned in our work. They correspond to the row and column numbers of the stated terms in the 18,000 x 18,000 co-occurrence count matrix.*

*This profile, for the record, is obtained using the normalization $A_{ab} = \frac{fab}{fa \cdot fb}$, over the highest frequency terms, using the 100 largest associates in each row, approximately. This defines the matrix used, and we note that only first-generation associations are involved here.

Each column of A can be regarded as the profile of the corresponding terms. The output shown in figure 4 can now be defined as follows. First describe the recognized input terms as a vector q: an 18,000 dimensional column vector showing 1's in the positions asterisked in figure 4 (i.e., in rows 9059, 11,183, etc.) and zeros elsewhere.

The weights in figure 4 are a scalar times the result of the matrix multiplication w = Aq. This profile is thus the sum of the profiles for MAN, PARTIAL, UNIT, PROCEDURE, etc.

BATCH 6 ASSOCIATIVE RETRIEVAL PAGE 1

WEIGHT	WORD	TERM NUMBER
.3768+	MAN	9059*
.3217+	PARTIAL	11183*
.3195+	UNIT	17197*
.3037+	PROCEDURE	12359*
.3026+	ZERO	18226*
.2812+	SUPPORT	15744*
.2767+	ACTIVATION	123*
.2418+	RING	13508*
.2227+	GRAVITY	6187*
.2115+	THREE	16421*
.1361+	SIMULATION	14425*
.1271+	WEIGHTLESSNESS	17855
.1170+	MOTOR	9919*
.0821+	MANNED SPACE FLIGHT	9049
.0753+	LIFE	8352
.0742+	MANNED SPACECRAFT	9047
.0742+	ORDER	10860
.0697+	ASTRONAUT	940
.0675+	CASE	2154
.0652+	DIMENSIONAL	3882
.0618+	PHYSIOLOGICAL RESPONSE	11646
.0585+	MERCURY PROJECT	9298
.0585+	SPACE FLIGHT	14894
.0585+	LOAD	8537*
.0562+	AEROSPACE MEDICINE	241
.0529+	SPACE ENVIRONMENT	14890
.0517+	APOLLO PROJECT	759
.0517+	MINUTEMAN ICBM	9644
.0506+	PSYCHOLOGY	12533
.0495+	METHOD	9437*
.0484+	MANNED	9046
.0484+	MAINTENANCE	8973
.0483+	SUBSYSTEM	15613
.0461+	PHYSIOLOGY	11648
.0461+	SIMULATOR	14426
.0461+	SYSTEM	15932*
.0438+	DIFFERENTIAL	3813
.0438+	MEMORY	9274
.0427+	TRAINING	16690
.0427+	DERIVATIVE	3673
.0427+	COMMAND	2784
.0427+	HUMAN	6885
.0405+	CLOSED	2610
.0405+	MACHINE	8823
.0405+	GROUND	6227
.0394+	RENDEZVOUS	13297
.0393+	TOLERANCE	16577
.0393+	ELLIPSE	4647

FIGURE 4 ASSOCIATION PROFILE FOR THE FULL TEXT REQUEST IN FIGURE 1

WEIGHT	WORD	TERM NUMBER
.0360+	ENVIRONMENT	4789
.0360+	ATTITUDE	1058
.0360+	CYLINDRICAL SHELL	3416
.0360+	TORQUE	16613
.0360+	GRAVITATIONAL	6183
.0349+	DIFFERENTIAL EQUATION	3809
.0348+	BUCKLING	1892
.0337+	LABORATORY	8100
.0326+	SPACE	14918*
.0326+	SPACE SCIENCE	14919
.0326+	WINDING	17936
.0326+	FIRING	5318
.0315+	HANDLING	6362
.0315+	ATTITUDE CONTROL	1055
.0315+	AEROSPACE	242
.0315+	SECOND	14002
.0315+	SOLID PROPELLANT ROCKET ENGINE	14757
.0304+	SPACECRAFT POWER SUPPLY	14854
.0303+	RELATIVITY	13262
.0303+	FACILITY	5111
.0292+	ANIMAL STUDY	640
.0292+	ANALOG	589
.0292+	EXPLORATION	5025
.0292+	WEAPON	17813
.0281+	ADAPTATION	138
.0281+	MOON	9874
.0281+	MEDICINE	9253
.0270+	RESPIRATION	13367
.0270+	COMMUNICATION SYSTEM	2798
.0270+	MANAGEMENT	9011
.0270+	METEOROID	9409
.0270+	MISSION	9689
.0270+	CENTER	2248
.0270+	SPACE PROGRAM	14913
.0259+	HYBRID	6914
.0258+	DIFFERENCE	3805
.0258+	HUMAN PERFORMANCE	6883
.0247+	RELIABILITY	13276
.0247+	ANIMAL	639
.0247+	PILOT	11689
.0247+	LOADING	8533
.0236+	REVOLUTION	13429
.0236+	SATELLITE ORBIT	13806
.0236+	NASA PROGRAM	10055
.0236+	OPERATOR	10762
.0236+	EXTRATERRESTRIAL	5083
.0236+	REQUIREMENT	13317
.0236+	OPTIMAL CONTROL	10817

FIGURE 4 (CONTINUED)

5. Comparison of the Association Profile and the Analyst's Search

i) Successful Discovery of Alternate Vocabulary

The association process identifies the term WEIGHTLESSNESS as the most closely associated term that was not present in the request received. The prominence given to WEIGHTLESSNESS in the profile should be construed as a suggestion that the requestor consider using this terminology as alternate vocabulary in phrasing a request to the NASA collection. The requestor did not use this term in his full text request. (We speculate that the term did not enter his mind as a way of describing what was wanted.) He used, rather, the terminology "zero gravity" and "partial gravity" to express his idea. (The analyst did use the term WEIGHTLESSNESS in his search prescription.) We believe it is likely that a requestor (were he to receive this profile as a suggestion list after submitting his request) would recognize WEIGHTLESSNESS to be a term he ought to use in his search.

ii) Other Means of Vocabulary Expansion

Naturally, a requestor who wishes to find good search terms could use available printed term lists and thesaurus-like guides.

This same example illustrates some of the encumberances of doing this, and we include our observations for completeness. Let us suppose the requestor had tried to find alternate terminology for ZERO GRAVITY using the Guide to the Subject Indexes for STAR (April 1964 or February 1965 version). He would not have been led from ZERO GRAVITY to WEIGHTLESSNESS as was done with the aid of the association list. Certainly the suggestive cross-reference could have been placed in a printed thesaurus, but it wasn't. In the EJC Thesaurus, moreover, an attempt to use ZERO GRAVITY leads to finding no entry. It is not a recognized heading. So one tries GRAVITY (looking for zero-gravity simulators and the like) and finds:

GRAVITY CONCENTRATORS
GRAVITY CONVEYORS
GRAVITY DAMS
GRAVITY DRAINAGE (RESERVOIRS)
GRAVITY METERS

each with a set of irrelevant terms. Having determined that GRAVITY SIMULATION is not a heading, one turns to

GRAVITY
  USE GRAVITATION.

Turning the page, an alphabetic scan locates this new term, and under GRAVITATION one finally finds (alphabetically last) the first acceptable alternate term -- the term WEIGHTLESSNESS itself. This is work, the kind

of work which is an obstacle to active investigation of the literature by interested users.

Use of the association program is a far less arduous and (in our judgment) an effective way to home in on the right vocabulary to use in formulating a search.

iii) The Analyst's Term "Weightlessness Simulation"

Ideally, in the present search, the requestor should be led to the term WEIGHTLESSNESS SIMULATION for use in a search conducted through the NASA collection as a whole. In the pre-1965 subset, this term occurred only four times; it is not one of the terms in the association matrix and hence is not retrieved in the association profile. However, the requestor is easily led to this term once WEIGHTLESSNESS has been identified as relevant terminology: a display of alphabetically similar terms would lead him there at once.

Once the requestor arrives at the term WEIGHTLESSNESS SIMULATION, we consider him "home." The analyst wanted to see every document indexed by this tag that was within the subcollection he searched, and we concur. Thus, the request we have discussed exemplifies the way an association profile can guide an untrained requestor quite directly to one or more multiple-word terms (if there are any) which obviously are descriptive of what he wants. We have discussed this process at length (in Reference 12), and this demonstration is consistent with those results.

6. Discussion of Other Terms on the Profile. - Considering the austere circumstances (no pre-editing of any consequence) under which this request was run, it is not uninteresting to examine some of the other terms on the list. (We ignore the starred terms at the top of the profile -- they were in the request, and the machine process is currently tuned to place request terms high on the list.) We treat the terms in the ranked order in which they appear on the profile.

MANNED SPACE FLIGHT is an idea that is implicit in the request. While it is far more general a topic than the requestor intended, the term is a reasonable associate for a statement that deals with a man in a space suit in a weightless environment.

LIFE and MANNED SPACECRAFT are judged to be derivative ideas from the one above.

ORDER is an artifact,* derived, we believe, from the ideas of PARTIAL ORDER, ZERO ORDER, and THREE ORDER.

---

*Artifacts are to be expected in profiles generated, as these were, without careful "tuning" of the procedure to yield the best achievable results. It is known that NASA indexers systematically include the important constituent words of a published multiword index term in the index set; we have not yet determined the best adjustment for this, but see no impediment to doing so.

ASTRONAUT has been explained.

CASE comes from MOTOR CASE, we believe, and is an artifact.

DIMENSIONAL probably comes from THREE DIMENSIONAL and is also an artifact.

The next group of terms, save for the very unhelpful request terms LOAD, METHOD and SYSTEM -- note that they were demoted by the machine-generated profile -- contains a number of plausibly related ideas.

- ✓ PHYSIOLOGICAL REPONSE
- ✓ MERCURY PROJECT
- SPACE FLIGHT
- ✓ AEROSPACE MEDICINE
- ✓ SPACE ENVIRONMENT
- APOLLO PROJECT
- MINUTEMAN ICBM
- ✓ PSYCHOLOGY
- MANNED
- MAINTENANCE
- SUBSYSTEM
- ✓ PHYSIOLOGY

It is appealing to explain the presence of the checked terms (AEROSPACE MEDICINE, PHYSIOLOGICAL RESPONSE, etc.) as a reflection of the deep concern (in the early days of the space program) with which the effects of weightlessness on man were approached. When we pose the given request to the collection we processed -- the old NASA segment -- it is possible to conjecture that the profile is suggesting that the early collection deals heavily with this aspect of the request. We have no evidence, however, that the collection is so slanted, even though it does seem likely.

There is little more to be gained from conveying an appreciation of just one profile. Nevertheless, having proceeded through the list to the 20th associated term, we note that the term SIMULATOR is also suggested: another important term in the analyst's view, one we have spoken about before as carrying the notion of "devices" in the analyst's prescription.

Conclusion: An association profile for an unedited full-text request, employing only the individual single words in the text of that request can (and in our experience almost invariably does) lead the requestor to a high precision search if one exists.

Comment: There is considerable room for improvement in the quality of the machine output when it is intended for human inspection and interpretation of the kind we just performed. Association profiles like these have not yet been "human engineered" for use either (a) as vehicles for making positive suggestions as to alternate vocabulary or (b) as devices for reporting back how the searching system is going to interpret the given request. As the associative information comes to be incorporated into machines with more significant capabilities than the 1401, it will be practical to pay more attention to providing more direct and more concise reports to the requestor.

D. Associative Search on the Pre-1965 Collection

A demonstration search, conducted principally to assure NASA that the programs operate without fault, was performed through the entire set of 95,509 documents contained in the version of the Linear File which we received. This search was conducted using one of 20 searches which had recently been conducted by the NASA Facility; because the output from that search had been evaluated by the analyst it was possible to compare the two sets of retrieved documents. It was important to insure that the associative search would retrieve the same documents -- to demonstrate that no serious errors were present in the programs or in our processing of the data from the collection.

1. Choice of a Search. - Since no particular hypothesis was under test, we chose the search prescription to be tested without particular deliberation. We were, however, guided by some general considerations.

We have learned during this work with the NASA collection that there are quite a large number of subject areas included in it whose terminology is entirely too technical or too obscure to permit non-experts to understand what is going on. In the area of Chemistry of Propellants, for instance, we found it impossible, as non-experts, to judge whether named compounds are or are not associated with the terms in a request. We ruled out using such technical questions and chose the search that seemed simplest to understand and might have the greatest general interest to the readers of the report. We also restricted ourselves to searches that had several terms in the association vocabulary of 1000 high-frequency terms.

The search we chose was entitled "Rendezvous and Docking" by the analyst who prepared it. The original full-text request received by the facility is shown in figure 5 and the search prescription appears in figure 6.

It is noteworthy that the librarian who prepared the full-text request in figure 5 was quite explicit as to the intended coverage and that she did employ quite a few NASA terms in the text of her letter. This particular example of a request does not look like the language of an unassisted technical person -- it seems to have been processed by a knowledgeable librarian before she sent it in.

2. The Analyst's Search Prescription. - In contrast to the previous illustration, the analyst's search prescription is the basic input this time. Though conducting an associative search is not expected to improve upon the search an expert analyst can conduct, use of his terms provided a positive check on whether we had accurately retained all the indexing which was present on the Linear File. Thus in the present example, the set of terms in the analyst's search prescription was transcribed, all were given equal weights, and the result was submitted to the entire associative processing operation using the Retrieval Programs.

June 10, 1966

Machine Search Branch
Reference Department
NASA Scientific & Technical Information
Facility
P. O. Box 33
College Park, Md. 20740
ATTN: Philip F. Eckert

Gentlemen:

I would like to request for a literature search done on Rendezvous and Docking including: spacecraft rendezvous, space rendezvous maneuver, orbital rendezvous maneuver, rendezvous guidance, rendezvous sensors, and rendezvous docking.

Our facility identification code number is 723, and we are requesting this service under contract. . . . Would you please consider this a RUSH request. Thank you.

Sincerely yours,

FIGURE 5 FULL-TEXT REQUEST ON "RENDEZVOUS AND DOCKING"

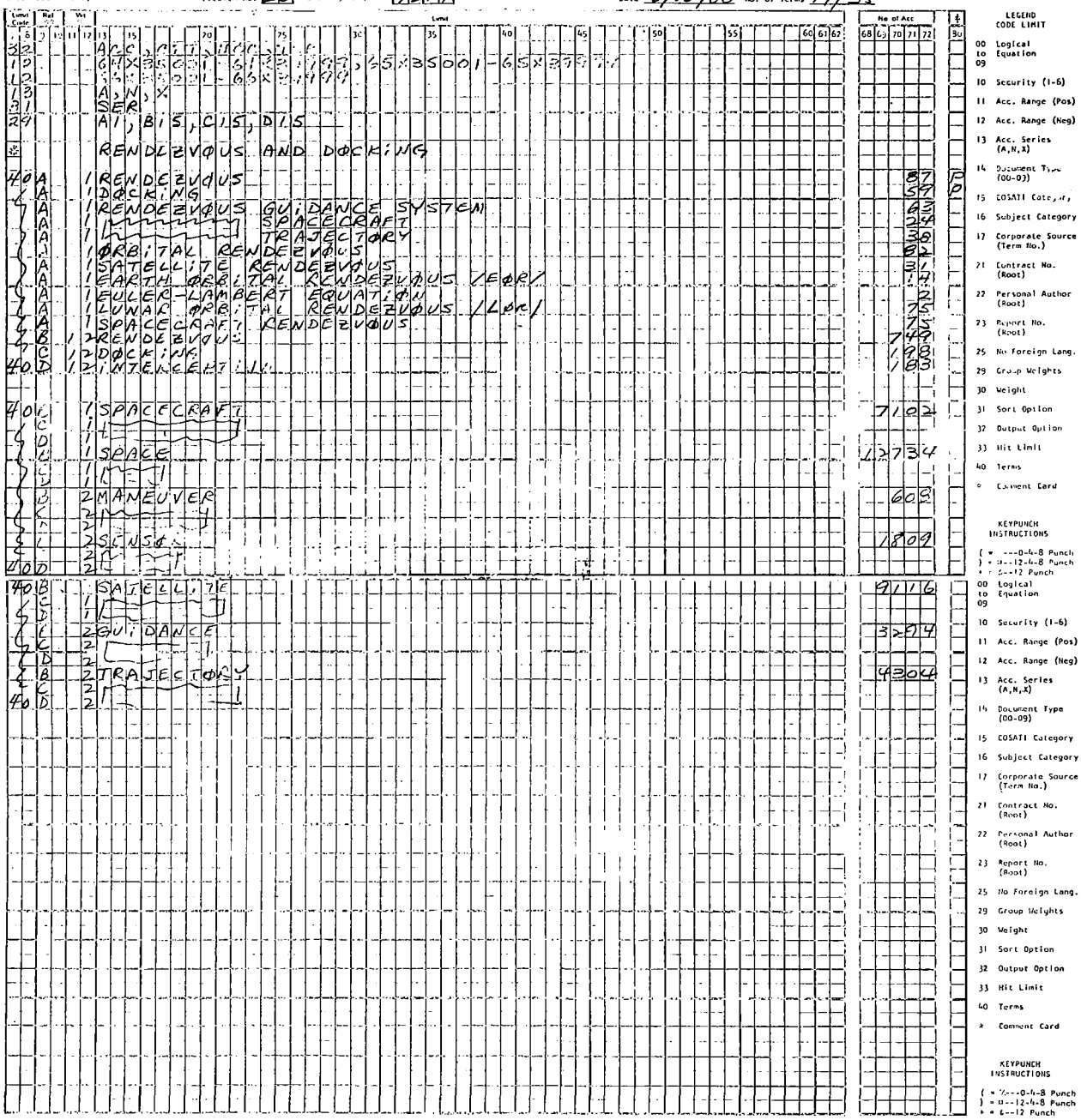


FIGURE 6 ANALYST'S SEARCH PRESCRIPTION FOR "RENDEZVOUS AND DOCKING"

3. Association Phase. - The input terms are displayed in figure 7 and figure 8 shows the association profile that was produced using the 1000 x 1000 association matrix over the high frequency terms. Of the request terms, i.e., the terms in the analyst's search prescription, only eight were among the set of 1000 highest frequency NASA terms. As we see in the profile in figure 8, the recognized terms in the input request (marked *) were:

MANEUVER
RENDEZVOUS
SENSOR
TRAJECTORY
SPACECRAFT
SATELLITE
SPACE

The remaining terms were not in the association vocabulary. The reader will recognize that this set of seven terms does not quite capture the content of the request. For example, the idea of docking is lost, and the term set is considerably less specific than the request was.

a. Personal Reaction to the Association Profile. - The profile in figure 8 was reviewed at the time the search was run -- that is, during the pause between the association phase and document searching phase when intervention is allowed. Our reaction was to find it unusually neutral: no term in the top 100 or so seemed to be especially applicable to the request as we understood it, nor did any one seem to be significantly inapplicable. The only term in the top 20 or so that caused a moment of hesitation was LAUNCH, but then we recalled that the time of launch was a critical factor in achieving rendezvous. This doubt was enough to cause us to leave it in, rather than attempting to select the "best" terms from the association profile.

Our inspection of the profile and the decision to leave it alone took about three minutes.

4. Difficulty of Reacting Accurately. - A real requestor, unlike those of us who are bystanders, has a very well-developed idea of what he is looking for. It is to be expected that he can select and reject terms from a profile somewhat more easily than other persons for he is motivated and has a strong opinion.

To illustrate this point, suppose we were interested only in rendezvous and docking operations in the vicinity of earth. Then it would be evident that terms like MOON, MARS, etc. could be deleted. Their deletion does not mean there will be nothing about MOON or MARS retrieved. It merely means that MOON and MARS documents will be less prominent on the retrieved list than they would otherwise be.

Some readers may find they have strong opinions toward the terms SPACECRAFT PROPULSION or REENTRY, feeling that these terms are not helpful in describing what they want to see. The printing of an association profile is done precisely so that such views can be communicated to the search system, and the requestor's choice among terms in the profile is a fine adjustment to the "normal" output of the associative search. In our experience, the best action for the requestor to take in editing a profile is to delete unwanted terms without revising the weights assigned (12).

NO. 2692 SEARCH PRESCRIPTION--RENDEZVOUS

THRESHOLD .000010

TERM	INPUT WEIGHT	NORMALIZED
RENDEZVOUS	9999	2293
DOCKING	9999	2293
RENDEZVOUS GUIDANCE SYSTEM	9999	2293
RENDEZVOUS SPACECRAFT	9999	2293
RENDEZVOUS TRAJECTORY	9999	2293
ORBITAL RENDEZVOUS	9999	2293
SATELLITE RENDEZVOUS	9999	2293
EARTH ORBITAL RENDEZVOUS /EOR/	9999	2293
EULER-LAMBERT EQUATION	9999	2293
LUNAR ORBITAL RENDEZVOUS /LOR/	9999	2293
SPACECRAFT RENDEZVOUS	9999	2293
INTERCEPTION	9999	2293
SPACECRAFT	9999	2293
SPACE	9999	2293
MANEUVER	9999	2293
SENSOR	9999	2293
SATELLITE	9999	2293
GUIDANCE	9999	2293
TRAJECTORY	9999	2293

FIGURE 7 INPUT TO ASSOCIATION PHASE - ANALYST'S "RENDEZVOUS AND DOCKING" TERMS

WEIGHT	WORD	TERM NUMBER
.8965+	MANEUVER	9017*
.7681+	RENDEZVOUS	13297*
.3531+	SENSOR	14116*
.2958+	GUIDANCE	6258*
.2040+	TRAJECTORY	16694*
.1673+	MANNED SPACECRAFT	9047
.1605+	SPACECRAFT	14861*
.1444+	APOLLO PROJECT	759
.1421+	NAVIGATION	10077
.1329+	MISSION	9689
.1284+	MANNED	9046
.1238+	ORBIT	10857
.1215+	LANDING	8168
.1215+	TERMINAL	16148
.1192+	ATTITUDE	1058
.1054+	SPACE FLIGHT	14894
.1054+	ATTITUDE CONTROL	1055
.0963+	MOON	9874
.0963+	SPACECRAFT PROPULSION	14855
.0894+	MANNED SPACE FLIGHT	9049
.0871+	COMMAND	2784
.0848+	SPACE VEHICLE	14935
.0825+	SPACE PROGRAM	14913
.0802+	LAUNCH	8235
.0802+	SATELLITE	13812*
.0802+	INTERPLANETARY	7569
.0756+	SATELLITE ORBIT	13806
.0756+	ASTRONAUT	940
.0756+	ENTRY	4779
.0733+	MARS	9126
.0688+	TRACKING	16659
.0665+	PAYLOAD	11244
.0665+	SPACE	14918*
.0665+	CORRECTION	3092
.0664+	SIMULATOR	14426
.0642+	SPACECRAFT POWER SUPPLY	14854
.0642+	SPACE SCIENCE	14919
.0642+	REENTRY	13182
.0642+	REQUIREMENT	13317
.0619+	BOOSTER	1709
.0619+	PROPULSION	12458
.0619+	TARGET	16013
.0596+	VEHICLE	17407
.0596+	EXPLORATION	5025
.0573+	MAN	9059
.0573+	COMMUNICATIONS SATELLITE	2802
.0573+	MERCURY PROJECT	9298
.0550+	MARS /PLANET/	9127

FIGURE 8 ASSOCIATION PROFILE FOR ANALYST'S TERMS IN FIGURE 7

WEIGHT	WORD	TERM NUMBER
.0527+	NASA PROGRAM	10055
.0527+	SATELLITE OBSERVATION	13803
.0527+	STATION	15268
.0527+	THRUST	16453
.0527+	SUBSYSTEM	15613
.0504+	PILOT	11689
.0504+	INERTIA	7341
.0504+	DISPLAY	3980
.0481+	PLANET	11789
.0481+	POSITION	12128
.0458+	METEOROID	9409
.0458+	LAUNCH VEHICLE	8239
.0458+	GYROSCOPE	6300
.0458+	TELEVISION	16082
.0435+	SPACE ENVIRONMENT	14890
.0435+	STABILIZATION	15167
.0435+	TELEMETRY	16071
.0412+	LUNAR	8770
.0412+	CELESTIAL	2231
.0412+	WEIGHTLESSNESS	17855
.0412+	REFERENCE	13187
.0412+	FLIGHT	5448
.0389+	PLANETARY	11782
.0389+	RECOVERY	13131
.0389+	FLIGHT TEST	5456
.0389+	ELECTRIC PROPULSION	4427
.0389+	INSTRUMENTATION	7460
.0366+	SUPPLY	15742
.0366+	VENUS	17434
.0366+	CAPABILITY	2046
.0344+	SCIENCE	13918
.0344+	ELLIPSE	4647
.0344+	GRAVITATIONAL	6183
.0343+	ACQUISITION	105
.0321+	REENTRY VEHICLE	13185
.0321+	EARTH	4254
.0321+	VISUAL	17591
.0321+	ORIENTATION	10896
.0321+	ALTITUDE	498
.0321+	BALLISTICS	1242
.0298+	OPTIMIZATION	10819
.0298+	ACCURACY	62
.0275+	LIFT	8369
.0275+	PATH	11233
.0275+	PROGRAMMING	12385
.0275+	OPTIMUM	10821
.0275+	ARTIFICIAL	873
.0275+	CONTROL	3010

FIGURE 8 (CONTINUED)

5. Submission of the Phase II Profile Search. - Figure 9 shows the profile submitted to the Phase II program for the search conducted through the pre-1965 collection. At the head of the list appear the 11 terms in the analyst's search prescription which were not in the 1000-term association vocabulary. These were all appended to the profile with equal weights. However, we made these weights lower than they would normally be made for one to use in a real search since there was an interest in seeing the effect of the association weights upon the document ranking produced.

The result, then, is a slight modification of the "fully automatic associative search" option, wherein a request is submitted to the program and the only output the requestor sees is the final listing of retrieved documents. For in this example, no selection of "preferred" search terms was made by us.

Figure 10 shows the first few documents retrieved -- those ranked highest by the associative search process. Rather than presenting the whole list for the reader's inspection, we turn to comparing the associative search output with what the analyst had earlier chosen to be relevant and summarize the quality of the search using his judgments.

6. A Brief Appraisal of the Retrieved Documents. - We have appraised the output of the associative search in two ways. First, we conducted a blind evaluation -- judging the output documents for relevance without knowledge of the analyst's search results. We then obtained the analyst's results and made those comparisons of the two outputs that were directly obtainable from the data. These two appraisals are summarized in this section, and they show that the two searches were closely comparable in performance (as they were meant to be).

a. Our own appraisal. - The associative search output contained 437 documents whose notations of content (NOC) were printed, in ranked order, as they would be presented to the requestor. This set of 437 retrieved documents included 126 classified documents. The whole list (both classified and unclassified documents) was presented to a single judge who assessed the documents for relevance to the request.

The judge began his evaluation with instructions to assess relevance on a four point scale. But he found such detailed judgment too difficult to accomplish using the NOC. (This judge had previously had no difficulty evaluating 50-100 word abstracts on a four point scale in other experiments using another collection.) The judge felt that not enough information is given in the NOC to permit him to be comfortable with his judgments. He felt he was being arbitrary and inconsistent when he tried to make a distinction between "peripherally relevant" and "quite relevant" using only the 10 or 12 words of text that comprise the NOC. The point scale was therefore abandoned after fewer than 50 documents had been looked at, and a binary scale (1 = relevant enough to be sent to the requestor, 0 = apparently irrelevant -- don't send to the requestor) was used.

THRESHOLD 06000
CARD INPUT LIMIT IS 100 OUTPUT LIMIT IS 100

TERM	INPUT WEIGHT
DOCKING	9999
RENDEZVOUS GUIDANCE SYSTEM	9999
RENDEZVOUS SPACECRAFT	9999
RENDEZVOUS TRAJECTORY	9999
ORBITAL RENDEZVOUS	9999
SATELLITE RENDEZVOUS	9999
EARTH ORBITAL RENDEZVOUS /EOR/	9999
EULER-LAMBERT EQUATION	9999
LUNAR ORBITAL RENDEZVOUS /LOR/	9999
SPACECRAFT RENDEZVOUS	9999
INTERCEPTION	9999
MANEUVER	8965
RENDEZVOUS	7681
SENSOR	3531
GUIDANCE	2958
TRAJECTORY	2040
MANNED SPACECRAFT	1673
SPACECRAFT	1605
APOLLO PROJECT	1444
NAVIGATION	1421
MISSION	1329
MANNED	1284
ORBIT	1238
LANDING	1215
TERMINAL	1215
ATTITUDE	1192
SPACE FLIGHT	1054
ATTITUDE CONTROL	1054
MOON	0963
SPACECRAFT PROPULSION	0963
MANNED SPACE FLIGHT	0894
COMMAND	0871
SPACE VEHICLE	0848
SPACE PROGRAM	0825
LAUNCH	0802
SATELLITE	0802
INTERPLANETARY	0802
SATELLITE ORBIT	0756
ASTRONAUT	0756
ENTRY	0756
MARS	0733
TRACKING	0688
PAYLOAD	0665
SPACE	0665
CORRECTION	0665
SIMULATOR	0664
SPACECRAFT POWER SUPPLY	0642
SPACE SCIENCE	0642

FIGURE 9 PROFILE USED FOR ASSOCIATIVE SEARCH ON "RENDEZVOUS AND DOCKING"

BATCH 1 CONTINUED.

TERM	INPUT WEIGHT
REENTRY	0642
REQUIREMENT	0642
BOOSTER	0619
PROPULSION	0619
TARGET	0619
VEHICLE	0596
EXPLORATION	0596
MAN	0573
COMMUNICATIONS SATELLITE	0573
MERCURY PROJECT	0573
MARS /PLANET/	0550
NASA PROGRAM	0527
SATELLITE OBSERVATION	0527
STATION	0527
THRUST	0527
SUBSYSTEM	0527
PILOT	0504
INERTIA	0504
DISPLAY	0504
PLANET	0481
POSITION	0481
METEOROID	0458
LAUNCH VEHICLE	0458
GYROSCOPE	0458
TELEVISION	0458
SPACE ENVIRONMENT	0435
STABILIZATION	0435
TELEMETRY	0435
LUNAR	0412
CELESTIAL	0412
WEIGHTLESSNESS	0412
REFERENCE	0412
FLIGHT	0412
PLANETARY	0389
RECOVERY	0389
FLIGHT TEST	0389
ELECTRIC PROPULSION	0389
INSTRUMENTATION	0389
SUPPLY	0366
VENUS	0366
CAPABILITY	0366
SCIENCE	0344
ELLIPSE	0344
GRAVITATIONAL	0344
ACQUISITION	0343
REENTRY VEHICLE	0321
EARTH	0321
VISUAL	0321
ORIENTATION	0321
ALTITUDE	0321

FIGURE 9 (CONTINUED)

A63-15841 GUIDANCE TECHNIQUES TO ACHIEVE INTERCEPTION OR RENDEZVOUS WITH AN EARTH SATELLITE AND COMPARISON WITH THOSE USED IN FIRE-CONTROL APPLICATIONS
A63-15841 SATELLITE INTERCEPTION WITH RENDEZVOUS. NORMAN E. SEARS AND PHILIP G. FELLEMAN/MASSACHUSETTS INSTITUTE OF TECHNOLOGY, INSTRUMENTATION LABORATORY, CAMBRIDGE, MASS./ IN-AIR, SPACE, AND INSTRUMENTS - DRAPER ANNIVERSARY VOLUME. NEW YORK, MCGRAW-HILL BOOK CO., INC., 1963, P. 120-137. 10 REFS.
FELLEMAN, P. G. SEARS, N. E.
14915 13037

A63-12388 DISCUSSION OF THE REASONS FOR THE NASA DECISION TO USE LUNAR ORBITAL RENDEZVOUS INSTEAD OF DIRECT FLIGHT OR EARTH RENDEZVOUS OPERATIONS
A63-12388 LUNAR RENDEZVOUS. JOHN C. HOUBOLT/NASA, LANGELY RESEARCH CENTER, HAMPTON, VA./INTERNATIONAL SCIENCE AND TECHNOLOGY, FEB. 1963, P. 62-70.
HOUBOLT, J. C.
11466 12587

N62-12834 NATIONAL AERONAUTICS AND SPACE ADMINISTRATION.
17728200LANGLEY RESEARCH CENTER, LANGELY STATION, VA.
ABORT TECHNIQUES FOR MANNED SPACECRAFT
N62-12834 NATIONAL AERONAUTICS AND SPACE ADMINISTRATION. LANGLEY RESEARCH CENTER, LANGLEY STATION, VA. SOME ABORT TECHNIQUES AND PROCEDURES FOR MANNED SPACECRAFT. JOHN M. EGGLESTON.)1962) 25 P. 12 REFS. FOR PRESENTATION AT THE NATIONAL IAS MEETING ON MAN*S PROGRESS IN THE CONQUEST OF SPACE, ST. LOUIS, MO., APR. 30-MAY 2, 1962. OTS- $2.60 PH, $0.95 MF.
EGGLESTON, J. M
2598 10280

A63-13960 DISCUSSION OF AN ORBITAL RENDEZVOUS BASE SYSTEM FOR ORBITAL ASSEMBLY AND LAUNCH OF MANNED LUNAR AND INTERPLANETARY VEHICLES
A63-13960 EARTH-LUNAR LOGISTICS EMPLOYING ORBITAL ASSEMBLY AND LAUNCH. N. V. PETERSEN, H. REICH AND R. S. SWANSON /NORTHROP CORP., NORTHROP SPACE LABORATORIES, HAWTHORNE, CALIF./ IN- SPACE LOGISTICS ENGINEERING. NEW YORK, JOHN WILEY AND SONS, INC., 1962, P. 339-433. 41 REFS.
PETERSEN, N. V. REICH, H.
13035 9674

FIRST FEW DOCUMENTS RETRIEVED IN ASSOCIATIVE SEARCH

(Figure 10)

Our judge found that 254 documents qualified as "relevant" in the search output. That is, he "accepted" them. Based on these judgments, we plotted the performance characteristic curve (12). In figure 11a, the plot shows cumulative relevance points -- in this case cumulative number of "accepted" documents -- vs. ranked on the output list. We see that by the time the judge had read 100 documents, he had found 92 accepted documents; by the time he had read 200 documents, he had found 162 accepted, etc.

Figure 11b is derived directly from the performance characteristic curve. It is a plot of the acceptance ratio as a function of the rank.

Before one can compute a single overall figure for acceptance ratio or precision ratio of a search with ranked output, it is necessary to determine how many documents are to be considered "retrieved". To this end, it is necessary to treat the evaluation in a more "dynamic" sense.

The documents are inspected, of course, in the order prescribed by the ranking, so that the curve, when read from left to right, is a track of the judge's reactions to the document NOC's he looked at. If we note, for example, the behavior of the curve between the 160th and 260th documents examined, we notice it has fairly steady slope of about .71. This means that during his inspection of this region of the output list, the judge was accepting better than two out of every three documents inspected. This acceptance rate changes, of course, as one proceeds from left to right, i.e., down the ranked list, as is clear from the plot of "local acceptance rate" in figure 11c. At the head of the list (the first 100 documents) the acceptance rate was better than nine out of ten, certainly a satisfactory rate. At the end of the list -- to the right of rank 360 on the plot -- the rate had dropped to less than two out of ten. This is considerable punishment with little reward.

Each person inspecting such a ranked list has his own tolerances for devoting effort to reading further on the list. If he is eagerly motivated to obtain an exhaustive search, he may proceed much further, continuing to look for a few more relevant ones even when he is finding only one relevant document in every ten. If he does not regard grinding through the list as distasteful, and his motivation is strong, he might proceed even further.

More typical behavior would be to feel that a point of diminishing returns has been reached at about the 260th ranked document, where the acceptance rate begins to deteriorate rapidly. Those who are irritated by finding more than one non-relevant document out of every ten inspected would be disenchanted at about the 100th ranked document. Those who want exactly 100 relevant documents would proceed to the 110th output document and stop there. Our judge went all the way through the output to define the curve's shape as far out as we thought was interesting. Hardly anyone would go so far in practice.

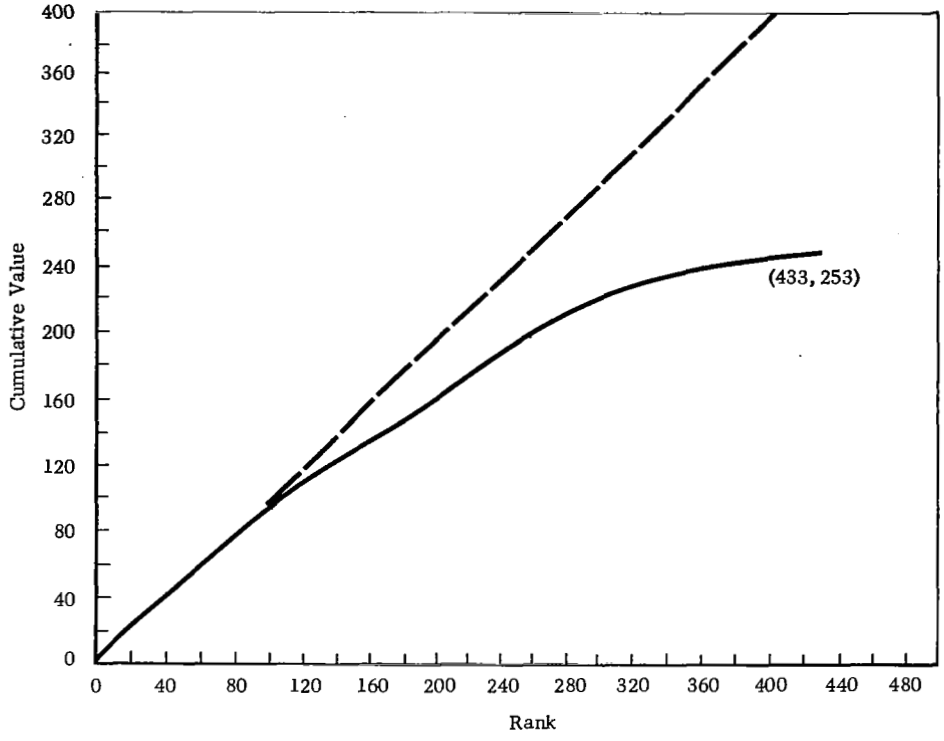


FIGURE 11A PERFORMANCE CHARACTERISTIC CURVE FOR THE ASSOCIATIVE SEARCH ON "RENDEZVOUS AND DOCKING"

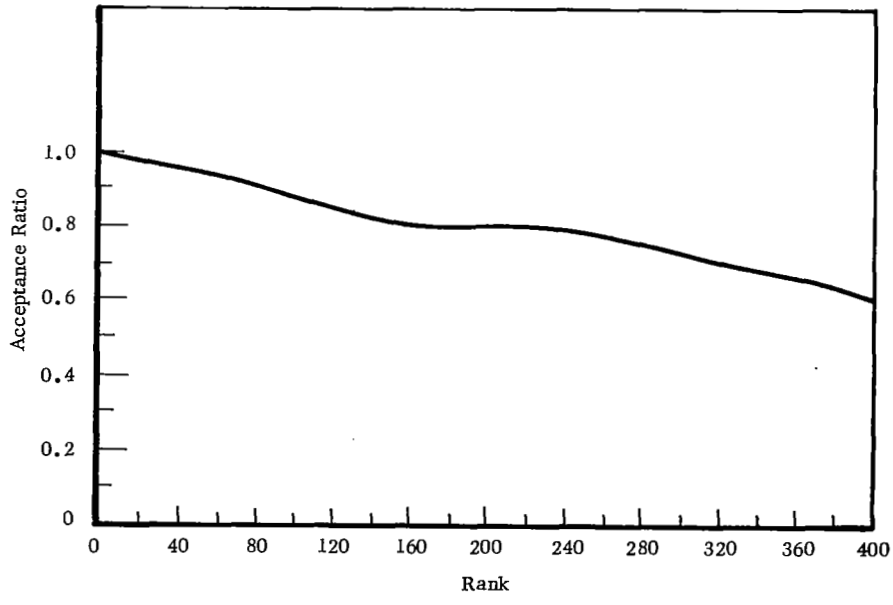


FIGURE 11B ACCEPTANCE RATIO AS A FUNCTION OF RANK

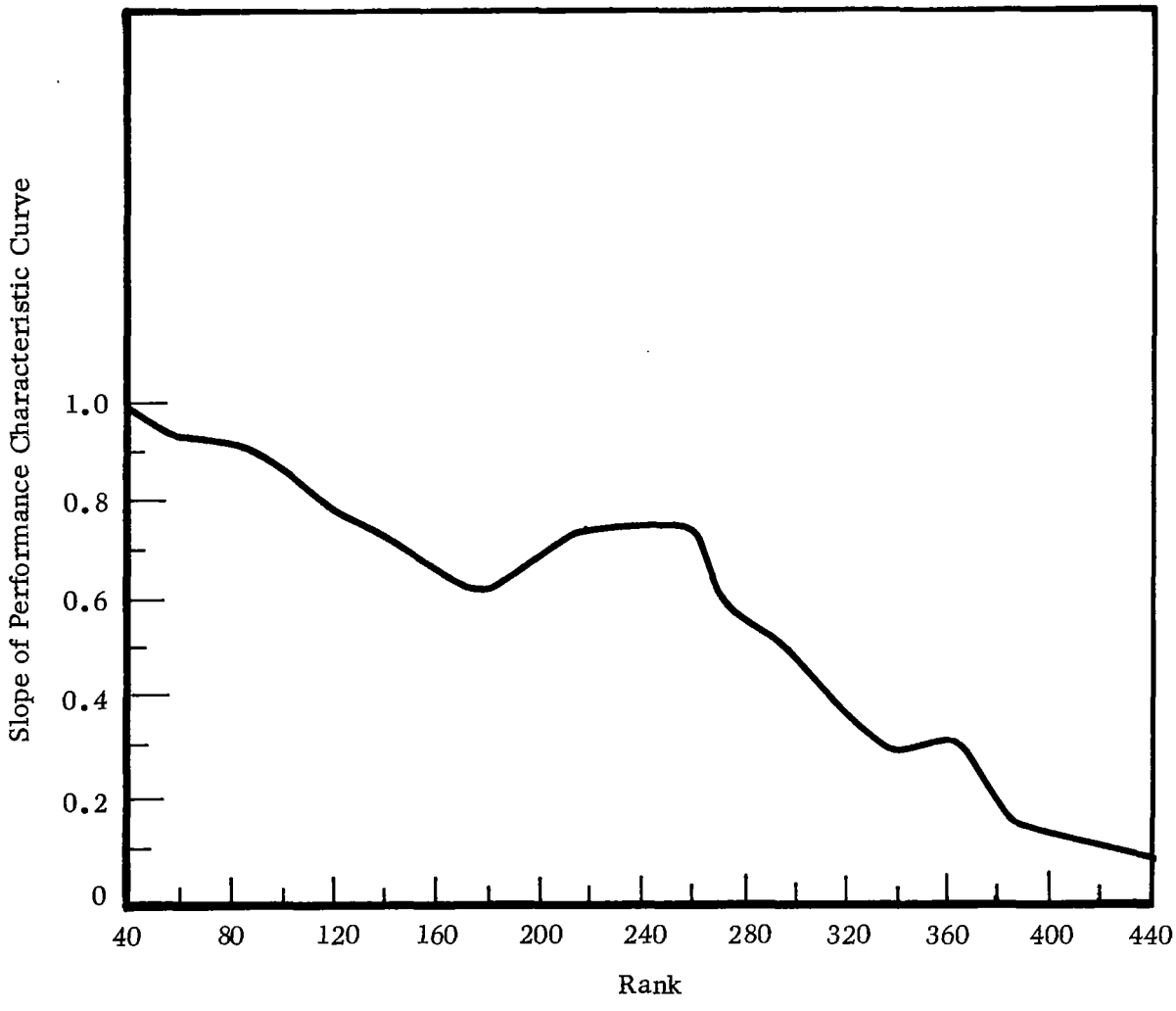


FIGURE 11C PLOT OF "LOCAL ACCEPTANCE RATE"

Some readers of this report may be willing to accept our judge's reactions as typical of their own. Such a reader can readily determine his own assessment of the quality of the output by studying the curves and estimating how far he would probably proceed down the list.

b. Comparison with the analyst's search. - We stated earlier that the analyst had conducted what he believed was a tight search and achieved an acceptance rate of 92%. We obtained the results of the analyst's search -- the set of unclassified documents which he accepted -- after completing our own evaluation in (a). We found he had accepted 179 documents that came from the same portion of the collection which we searched. (His output also contained additional accepted documents from the 1965-to-date portion of the NASA collection.)

Of the 179 documents, all but 36 were retrieved by the associative search.

A detailed analysis of the omitted documents was reported separately. The important points for the present discussion are that the majority of them would have been retrieved by the associative search had we given the extra emphasis to the analyst's terms discussed earlier.

The remaining 143 "overlap" documents (those which the analyst accepted and which we retrieved) had, as noted in (a) above, been previously evaluated for relevance by our judge. There were only 13 cases of disagreement -- i.e., our judge had earlier rejected 13 of the analyst's accepted documents. We have reviewed these discrepancies and most of them are reasonably explained by the difference between what the NOC states and the more complete information revealed about the document through the set of terms assigned. There are at most four documents where our judge would want the analyst to reconsider his appraisal. Our judge and the analyst agreed very well on the "overlap" documents -- over 90% agreement, even using different document descriptions.

Our judge, however, found 61 other unclassified documents in the associative search output which he judged relevant. We do not know whether the analyst would have agreed. We can deduce (because the analyst had a 92% acceptance ratio) that he probably only rejected about 15 documents in the set from which he accepted 179. He probably did not look at most of the 61 extra documents the associative search retrieved.

Conclusion: The fact that there were these 61 additional, allegedly relevant, documents in the associative search output merely demonstrates that the associative search seems to be accomplishing what it is meant to do. It in no way demonstrates the "superiority" of associative search over the analyst's performance. For example, had the analyst conducted a looser search, maybe he would have retrieved these 61 and more.

E. Second-Generation Association Profiles

1. Fundamentals. - The formula we have employed to measure the degree of association between two terms is $A_{ab} = \frac{fab}{fa \cdot fb}$, where fab is the co-occurrence count and fa, fb are the term frequencies. If fab = 0, i.e., the two terms do not co-occur, then the association strength between them is obviously zero. The measure is strongly dependent upon the co-occurrence count, and a basis for its use is readily seen. Let d be the total number of documents in the collection. Then fa/d is the probability that a document (picked at random from the collection) is posted to term a. Similarly fb/d is the probability that a document is posted to term b. The probability that, by chance, the document is posted to both a and b is the product of these probabilities, viz.:

$$P_{chance}\ (a \cdot b) = \frac{fa}{d} \cdot \frac{fb}{d}$$

The expected number of documents posted to both terms is simply d times the probability $P_{chance}\ (a \cdot b)$. The ratio below is readily computed:

$$\frac{\text{Observed No. of co-occurrences}}{\text{Expected No. of co-occurrences}} = \frac{fab}{\frac{fa}{d} \cdot \frac{fb}{d} \cdot d} = \frac{fab}{fa \cdot fb} \cdot d$$

Except for a constant of proportionality (d), this is the association measure A_{ab} we have used.

This is a "first-generation" association measure by virtue of being critically dependent on the co-occurrence of the two terms in question. If the two terms happen never to co-occur, then this formula effectively states they are not associated (or, more accurately, that evidence to that effect is not present in the data which was used in the model).

For some time, researchers have recognized that a simple co-occurrence-based association measure has this inherent limitation. Two synonymous terms, for instance, might well tend not to co-occur, yet they ought to be found "associated" by the statistical procedure. An ideal measure of the degree of association between terms should take such an effect into account.

Second-generation associations provide an appealing avenue toward circumventing this deficiency. The idea, intuitively, is this: if the first-generation association process is working fairly well, then the first-generation profiles for any term may be construed as a summary of the terminological environment of the term in question. It displays the co-occurring terms which are found, with surprising frequency, in the environment (i.e., an index set) containing the given term. The profiles of two nearly synonymous terms should resemble each other: AIRCRAFT's profile should resemble AIRPLANE's more closely than it resembles the profile of almost any other term.

While many computable definitions of "resemble" can be written, one resemblance measure with the virtue of simplicity is a straightforward correlation co-efficient: the profile (vector) for term a is correlated with the profiles (vectors) for all the other terms b, and the resulting set of correlation measures serves as a "second-generation association" profile. This class of procedures (measuring the resemblance of profiles) offers an avenue toward the machine discovery of synonyms and near-synonyms. Even though AIRCRAFT and AIRPLANE never co-occur, they could have a high degree of "second-generation association" as a consequence of the similarity of their first-generation profiles.

2. Practicality. - The practical problems involved in studying the similarity of profiles are quite serious when the scale of the data base under study is increased. The computational operations are extensive. First, one must obtain an association matrix over a sizeable segment of vocabulary employed in the data base. Each column of that matrix contains one term's associations with the rest of the vocabulary (the basic data used to create the term profiles we have exhibited so far).

One must then correlate that column with all the other columns of the association matrix to get one second-generation profile. The act of correlating just one column with the given one involves several hundred multiplications and additions. This yields a single number: the second-generation association strength between term a and term b. If there are 1000 terms in the subset of the vocabulary under study, the process needs to be repeated 1000 times to get the 1000 such numbers that make up one second-generation profile. To get the 1000 profiles that are available, it is necessary to repeat the foregoing package of steps 1000 times. If the basic correlation operation (the innermost loop of multiplication and addition plus "overhead" operations like testing for column no. matches, for overflow, etc.) can be accomplished in n hundredths of a second by computer, the total machine time involved for the whole task is one million times n.

Our tests indicate that n = 2 is an appropriate figure for the association matrices we have examined. If one were willing to throw away a good portion of the first generation association profile -- concentrating attention only on measuring resemblance using the "top few terms on the list" -- the number n would be reduced. But the scientific value of the results would be jeopardized. On the other hand, retaining and processing all the associations into which a given term enters -- each term is associated with almost all others -- is not necessary, and would raise the value of n considerably. The value n = 2 is a reasonable compromise, and it leads us to estimate that about six hours of computer time would be needed to calculate the whole set of second-generation association profiles. There are other costs as well -- other programs that need to be run. Thus, a figure in the vicinity of ten or twelve hours of 7094 time is a fair estimate of the cost of incorporating second-generation associations into a 1000 x 1000 matrix.

Forbidding though this investment may seem, it cannot be ruled out on a cost basis. It is an operation that needs to be done only rarely -- possibly every year or so when the association matrix is updated. If one can point to about $10,000 worth of improvement attributable to use of the second-generation associations (e.g., saving one-half man year of effort somewhere, or providing an improvement in performance or service that increases the number of active users of the NASA collection) then the cost is justified. In large operating systems such costs can be justified; research budgets, however, cannot embrace such investments easily.

3. Decisions Made. - In the belief that use of second-generation associations was both economically feasible and scientifically promising, we incorporated the capability to calculate and use them into the system design. Indeed, a reading of the section entitled "Matrix-Processing Program" in Part III will show fairly extensive capabilities for obtaining combinations of first, second, and higher generation associations as desired. In keeping with our objective of taking research ideas out of the laboratory, programs to incorporate such ideas into the Associative Retrieval System are fully operational.

But because the evidence as to the usefulness and need for second-generation associations remains ambiguous at the present time, we have not made the actual investment of incorporating the second-generation associations into the association matrix we have used in test operations. In our judgment it has been premature to do so. We believe further exploration and assessment of the pros and cons of such an action are required. Almost nothing is known about what the association matrix for an operating system of this size should "look like." It has seemed prudent to check first to see if the first-generation association matrix has the kind of deficiencies that could be corrected by using the second-generation associations rather than gambling on scanty evidence.

In lieu of a blanket investigation of second-generation association profiles, we have programmed, and used, a less efficient but far more flexible configuration of programs whose purpose is to permit the exploration of the contribution of second-generation associations. This arrangement involves using the 1401 Retrieval program twice, with some intermediate adjustments to obtain second-generation profiles whose mathematical properties are fully known. The procedure is developed in Appendix B where the mathematical correspondences are shown. For the present purposes, the details are not significant except to professional researchers in the field, and we skip them.

4. Some First-Generation Profiles for Synonym Pairs. - Before examining a second-generation profile, it is valuable to display, side-by-side, the first-generation profiles for several pairs of terms that could be thought of as near synonyms. We have included five such pairs from the 1000-term association vocabulary for which an association matrix was prepared. The profiles for these terms (over the 1000-term vocabulary) are shown in Figure 12:

ROCKET	-	MISSILE
FUEL	-	PROPELLANT
EXTRATERRESTRIAL	-	SPACE
MAN	-	HUMAN
VELOCITY	-	SPEED

WEIGHT	WORD
.2499+	ROCKET
.2199+	ROCKET NOZZLE
.1399+	SOLID PROPELLANT ROCKET ENGINE
.1299+	CASE
.1299+	SOUNDING
.1199+	MOTOR
.1099+	EXHAUST
.1099+	ROCKET ENGINE
.0899+	FIRING
.0799+	BOOSTER
.0799+	PROPELLANT
.0799+	MINUTEMAN ICBM
.0799+	NOZZLE
.0799+	ENGINE
.0799+	VALVE
.0699+	THRUST
.0699+	LIQUID PROPELLANT
.0699+	SOLID PROPELLANT
.0599+	IGNITION
.0599+	PROPULSION
.0599+	EROSION
.0500+	BALLISTICS
.0500+	SOLID
.0500+	STAGE
.0500+	CHAMBER
.0500+	PAYLOAD
.0400+	FUEL
.0400+	HYBRID
.0400+	REFRACTORY MATERIAL
.0400+	TANK
.0400+	LAUNCH
.0400+	LAUNCH VEHICLE
.0400+	SPACECRAFT PROPULSION
.0400+	SPECIFIC
.0400+	COMBUSTION
.0400+	IMPULSE
.0400+	INJECTION
.0400+	ION ENGINE
.0400+	OXIDIZER
.0400+	UPPER ATMOSPHERE
.0400+	UPPER
.0400+	VECTOR
.0300+	BALLOON
.0300+	ABLATION
.0300+	ALTITUDE
.0300+	GRAPHITE
.0300+	HYDRAZINE
.0300+	HIGH ALTITUDE
.6399+	MISSILE
.1599+	BALLISTICS
.1299+	WEAPON
.1199+	MINUTEMAN ICBM
.0799+	GUIDANCE
.0699+	FIRING
.0599+	RANGE
.0599+	TARGET
.0500+	TRACKING
.0500+	FLIGHT TEST
.0500+	LAUNCH
.0400+	BOOSTER
.0400+	ACQUISITION
.0400+	AERODYNAMICS
.0400+	HYDRAULICS
.0400+	PROJECTILE
.0400+	REENTRY
.0400+	MILITARY
.0400+	MOTOR
.0400+	SUBSYSTEM
.0400+	TRAJECTORY
.0400+	TERMINAL
.0400+	EXHAUST
.0400+	STAGE
.0400+	INERTIA
.0300+	ATTITUDE
.0300+	BLAST
.0300+	CAMERA
.0300+	AERODYNAMIC
.0300+	AIRBORNE
.0300+	GYROSCOPE
.0300+	PROJECT
.0300+	PROPULSION
.0300+	RADAR
.0300+	TELEMETRY
.0300+	TELEVISION
.0300+	FLEXIBILITY
.0300+	FLIGHT
.0300+	SPACE VEHICLE
.0300+	STABILIZATION
.0300+	COMMAND
.0300+	ENTRY
.0300+	ROCKET ENGINE
.0300+	SECOND
.0300+	ROCKET
.0300+	VALVE

END OF BATCH.

FIGURE 12a. FIRST GENERATION PROFILES FOR ROCKET AND MISSILE

WEIGHT	WORD	WEIGHT	WORD
.3799+	FUEL	.2499+	PROPELLANT
.3499+	FUEL CELL	.2399+	SOLID PROPELLANT
.1799+	LIQUID PROPELLANT	.2299+	LIQUID PROPELLANT
.1299+	URANIUM	.2299+	SOLID PROPELLANT ROCKET ENGINE
.1099+	OXIDIZER	.2099+	VALVE
.0999+	HYDROCARBON	.1799+	BINDER
.0999+	REACTOR	.1699+	PERCHLORATE
.0999+	TANK	.1499+	HYDRAZINE
.0999+	COOLANT	.1499+	OXIDIZER
.0999+	ELECTROLYTE	.1399+	AMMONIUM
.0899+	HYDRAZINE	.1399+	ROCKET ENGINE
.0899+	HYDRIDE	.1199+	HYBRID
.0899+	VALVE	.1199+	HYDRIDE
.0799+	FISSION	.1199+	SOLID
.0799+	SOLID PROPELLANT	.1199+	IGNITION
.0799+	CELL	.1099+	TANK
.0699+	BATTERY	.0999+	GRAIN
.0699+	BINDER	.0999+	MINUTEMAN ICBM
.0699+	MEMBRANE	.0999+	FIRING
.0699+	FLAME	.0999+	ROCKET NOZZLE
.0699+	COMBUSTION	.0899+	CASE
.0699+	IGNITION	.0899+	HIGH ENERGY
.0599+	ADDITIVE	.0899+	MOTOR
.0599+	HYBRID	.0899+	COMBUSTION
.0599+	POWERPLANT	.0899+	IMPULSE
.0599+	PROPELLANT	.0799+	ADDITIVE
.0599+	ROCKET ENGINE	.0799+	AMINE
.0500+	BIOCHEMISTRY	.0799+	FLUORINE
.0500+	LIQUID	.0799+	SPECIFIC
.0500+	SPACECRAFT POWER SUPPLY	.0799+	ROCKET
.0500+	SPECIFIC	.0699+	THRUST
.0500+	STORAGE	.0699+	EROSION
.0500+	CATALYST	.0699+	LIQUID
.0500+	CORE	.0699+	STORAGE
.0500+	ELEMENT	.0599+	BOOSTER
.0500+	ENERGY CONVERSION	.0599+	FUEL
.0500+	IMPULSE	.0599+	PROPULSION
.0500+	ROD	.0599+	NOZZLE
.0400+	HIGH ENERGY	.0599+	FLUORIDE
.0400+	HYDROGEN	.0599+	CHAMBER
.0400+	PRODUCT	.0599+	ENGINE
.0400+	RADIOACTIVITY	.0500+	EXHAUST
.0400+	PROPULSION	.0500+	EXPOSIVE
.0400+	MIXTURE	.0500+	STAGE
.0400+	NUCLEAR	.0500+	COMPOSITE
.0400+	THRUST	.0500+	CRYOGENICS
.0400+	CONTAMINATION	.0500+	INSTABILITY
.0400+	COOLING	.0500+	VISCOELASTICITY

FIGURE 12b. FIRST GENERATION PROFILES FOR FUEL AND PROPELLANT

WEIGHT	WORD	WEIGHT	WORD
.3769+	EXTRATERRESTRIAL	.1899+	SPACE
.0539+	LIFE	.1699+	SPACE SCIENCE
.0409+	MARS /PLANET/	.1599+	SPACE PROGRAM
.0319+	ORIGIN	.0999+	SPACE ENVIRONMENT
.0309+	MARS	.0899+	SPACECRAFT POWER SUPPLY
.0259+	BIOCHEMISTRY	.0899+	SCIENCE
.0249+	PLANET	.0799+	NASA PROGRAM
.0239+	METEORITE	.0799+	EXPLORATION
.0169+	EVOLUTION	.0699+	MANNED SPACE FLIGHT
.0159+	DUST	.0599+	ASTRONAUT
.0129+	BIOLOGY	.0599+	METEOROID
.0119+	EXPLORATION	.0599+	MANNED
.0109+	PLANETARY	.0599+	SPACE FLIGHT
.0099+	METABOLISM	.0599+	SPACE VEHICLE
.0099+	METEOR	.0599+	INTERPLANETARY
.0099+	ENVIRONMENT	.0500+	BIOLOGY
.0089+	SPACE SCIENCE	.0500+	AEROSPACE MEDICINE
.0089+	ORGANIC	.0500+	PROBE
.0079+	GALAXY	.0500+	SUPPLY
.0079+	RADIOACTIVITY	.0500+	MANNED SPACECRAFT
.0079+	DETECTION	.0500+	WEIGHTLESSNESS
.0079+	VENUS	.0400+	POWERPLANT
.0069+	SPACE ENVIRONMENT	.0400+	PROGRAM
.0069+	COSMIC	.0400+	MEDICINE
.0069+	INTERPLANETARY	.0400+	MERCURY PROJECT
.0059+	ASTRONOMY	.0400+	MISSION
.0059+	AEROSPACE MEDICINE	.0400+	TECHNOLOGY
.0059+	GEOLOGY	.0400+	ENVIRONMENT
.0059+	ISOTOPE	.0400+	EXTRATERRESTRIAL
.0050+	ANIMAL STUDY	.0400+	MAN
.0050+	PROBE	.0400+	SPACECRAFT PROPULSION
.0050+	STAR	.0400+	SPACECRAFT
.0050+	CELESTIAL	.0400+	STATION
.0050+	COMMUNICATION	.0400+	ELECTRIC PROPULSION
.0050+	SCIENCE	.0400+	LABORATORY
.0040+	BASE	.0400+	SIMULATOR
.0040+	ANIMAL	.0300+	ASTRONOMY
.0040+	MERCURY PROJECT	.0300+	AEROSPACE
.0040+	ENTRY	.0300+	APOLLO PROJECT
.0040+	EXPOSURE	.0300+	GALAXY
.0040+	SPACECRAFT PROPULSION	.0300+	HAZARD
.0040+	SPACE FLIGHT	.0300+	RELATIVITY
.0040+	SPACE	.0300+	MOON
.0040+	CONTAMINATION	.0300+	NAVIGATION
.0040+	DETECTOR	.0300+	SUN
.0040+	EARTH ATMOSPHERE	.0300+	SUPPORT
.0040+	DIOXIDE	.0300+	LIFE
.0040+	INFRARED RADIATION	.0300+	MARS

FIGURE 12c. FIRST GENERATION PROFILES FOR EXTRATERRESTRIAL AND SPACE

WEIGHT	WORD	WEIGHT	WORD
.6899+	HUMAN	.2969+	MAN
.5999+	HUMAN PERFORMANCE	.0359+	MANNED SPACE FLIGHT
.3599+	TOLERANCE	.0329+	MACHINE
.2599+	PSYCHOLOGY	.0219+	PSYCHOLOGY
.2199+	PHYSIOLOGICAL RESPONSE	.0199+	ASTRONAUT
.2099+	PERCEPTION	.0169+	MANNED SPACECRAFT
.1799+	PHYSIOLOGY	.0169+	WEIGHTLESSNESS
.1599+	AEROSPACE MEDICINE	.0159+	HUMAN
.1599+	FACTOR	.0159+	SPACE FLIGHT
.1599+	MAN	.0149+	MERCURY PROJECT
.1599+	WEIGHTLESSNESS	.0139+	PERCEPTION
.1299+	RESPIRATION	.0129+	TRAINING
.1199+	ASTRONAUT	.0119+	PHYSIOLOGY
.1199+	VISUAL	.0109+	HUMAN PERFORMANCE
.1099+	BEHAVIOR	.0099+	ANIMAL
.1099+	BLOOD	.0099+	APOLLO PROJECT
.1099+	MEDICINE	.0089+	MEDICINE
.1099+	METABOLISM	.0089+	TOLERANCE
.1099+	TRAINING	.0089+	RENDEZVOUS
.0999+	PILOT	.0089+	REQUIREMENT
.0899+	ANIMAL	.0089+	OPERATOR
.0899+	DISPLAY	.0089+	PHYSIOLOGICAL RESPONSE
.0799+	BIOCHEMISTRY	.0079+	BLOOD
.0799+	ACCELERATION	.0079+	CAPABILITY
.0799+	ADAPTATION	.0079+	AEROSPACE MEDICINE
.0799+	ENGINEERING	.0079+	SUPPORT
.0799+	OPERATOR	.0079+	EXPLORATION
.0799+	PERFORMANCE	.0079+	MAINTENANCE
.0699+	BIOLOGY	.0079+	COMMAND
.0699+	ANIMAL STUDY	.0079+	INFORMATION
.0599+	AEROSPACE	.0079+	PILOT
.0599+	MANNED SPACECRAFT	.0069+	BIOLOGY
.0599+	MANNED SPACE FLIGHT	.0069+	CAPSULE
.0599+	SPACE FLIGHT	.0069+	MANNED
.0599+	RESPONSE	.0069+	DISPLAY
.0599+	WORK	.0059+	ANIMAL STUDY
.0500+	AVIATION	.0059+	METABOLISM
.0500+	BODY	.0059+	SUBSYSTEM
.0500+	HAZARD	.0059+	EXPOSURE
.0500+	EXPOSURE	.0059+	FACTOR
.0500+	MACHINE	.0059+	LIFE
.0500+	MANNED	.0059+	SPACE SCIENCE
.0500+	CLOSED	.0059+	DURATION
.0500+	DOSAGE	.0059+	LANDING
.0500+	DURATION	.0059+	SELECTION
.0500+	INFORMATION	.0059+	VISUAL
.0500+	SELECTION	.0050+	AVIATION
.0400+	GRAVITY	.0050+	ADAPTATION

FIGURE 12d. FIRST GENERATION PROFILES FOR HUMAN AND MAN

WEIGHT	WORD	WEIGHT	WORD
.5699+	SPEED	.2399+	VELOCITY
.2799+	SUPERSONIC SPEED	.0599+	PROFILE
.1399+	SUBSONIC	.0599+	METER
.1299+	HYPERSONIC	.0599+	TURBULENT FLOW
.1099+	FLUTTER	.0500+	PROJECTILE
.0999+	AERODYNAMIC CHARACTERISTICS	.0500+	CHANNEL FLOW
.0899+	BLUNT BODY	.0500+	TURBULENCE
.0899+	HIGH SPEED	.0500+	TURBULENT
.0799+	BLUNTNESS	.0400+	POSITION
.0799+	SUPERSONIC	.0400+	RADIAL
.0799+	SWEEP	.0400+	STREAM
.0799+	DELTA	.0400+	FLOW FIELD
.0799+	WING	.0400+	COMPRESSIBILITY
.0699+	HYPERSONICS	.0400+	DETONATION
.0699+	TAKEOFF	.0400+	DRIFT
.0699+	LIFT	.0400+	INCOMPRESSIBILITY
.0699+	MACH NUMBER	.0400+	LAMINAR
.0599+	AERODYNAMIC	.0400+	LAMINAR BOUNDARY LAYER
.0599+	AERODYNAMICS	.0400+	WAKE
.0599+	AIRFOIL	.0300+	BOUNDARY LAYER
.0599+	ANGLE OF ATTACK	.0300+	BOUNDARY
.0599+	NOSE	.0300+	ACCELERATION
.0599+	LATERAL	.0300+	ANGLE
.0599+	LEADING	.0300+	AIRFOIL
.0599+	CHARACTERISTIC	.0300+	GAS FLOW
.0599+	DRAG	.0300+	GUN
.0599+	ROTOR	.0300+	HORIZONTAL
.0599+	WIND	.0300+	GRADIENT
.0599+	YAW	.0300+	HYPERVELOCITY
.0500+	TORQUE	.0300+	RADIUS
.0500+	LONGITUDE	.0300+	MEAN
.0500+	CONE	.0300+	METEOR
.0500+	DIRECTION	.0300+	MOTION
.0500+	TUNNEL	.0300+	MIXING
.0400+	BALLOON	.0300+	MOTION EQUATION
.0400+	BLADE	.0300+	TRAJECTORY
.0400+	ALTITUDE	.0300+	ENTRY
.0400+	HELICOPTER	.0300+	FRONT
.0400+	RATIO	.0300+	FLUCTUATION
.0400+	REENTRY VEHICLE	.0300+	FLOW
.0400+	SUPERSONIC FLOW	.0300+	FLAME
.0400+	ENTRY	.0300+	FRICTION
.0400+	FLIGHT TEST	.0300+	FLUID
.0400+	LOW	.0300+	LAYER
.0400+	SPHERE	.0300+	SONIC
.0400+	STATIC	.0300+	SOUND
.0400+	CONFIGURATION	.0300+	DOPPLER EFFECT
.0300+	BALLISTICS	.0300+	DUCT

FIGURE 12e. FIRST GENERATION PROFILES FOR SPEED AND VELOCITY

The overlap properties of these profiles are apparent by inspection, and there is evident resemblance in the profiles, as expected.

However, the most important result derived from these profiles is the observation that the words we thought of as synonyms* do co-occur. This is seen by the high position of synonym B on the first-generation profile for synonym A (and vice versa). While some collections may be so indexed that synonyms tend not to co-occur, this is apparently not the case in the NASA subcollection under study. Possibly this co-occurrence phenomenon is more prevalent among the high-frequency terms than it is among the lower-frequency ones; possibly it is more a property of "machine terms" than of "published terms"... These points have not been investigated at this writing.

Nevertheless, the observations that the pairs co-occur is evidence that second-generation associations are of marginal use in improving the 1000 x 1000 matrix we have studied. If the synonym is already associated by virtue of co-occurrence alone, the second-generation process will have a small effect (at best moving the synonym from 10th place to first place on the list) rather than a large one (like moving a term from 200th place to 10th place). While it would be aesthetically pleasing to find HUMAN at the head of the profile for MAN, its presence in 7th place is probably good enough for investigating practical uses of such profiles.

Accordingly, we have seen no pressing need to use second-generation associations in the 1000 x 1000 association matrix over the high-frequency terms. Nevertheless, we are continuing to explore the point further.

It is much more interesting to pursue a case where the two near-synonyms do not co-occur frequently, and we pursue this case in the next section.

5. Discussion of a Second-Generation Profile. - Figure 13 shows an association profile (over the 1000 high-frequency terms in the NASA collection) which incorporates second-generation effects. This profile is a combination of first- and second-generation associations. The terms labeled * have received weights that are affected by the combination step, that is, they are first-generation terms which have had more weight added by the second-generation process. We shall ignore these asterisked terms in this discussion, focusing attention on the unlabeled ones.

The unlabeled terms are second-generation associations. Although they may have co-occurred with the header terms a few times, that number of co-occurrences was so small that the degree of first-generation association was not considered significant. The unstarred terms, in short, were not considered associated in the first-generation computation. They are "pure" second-generation terms.

The term for which this second-generation profile was prepared is MAGNETOHYDRODYNAMIC FLOW. This term embraces phenomena encountered when an electrically conducting fluid is in motion in the presence of a magnetic field.

---

*VELOCITY - SPEED is the only exception so far noted among the high frequency pairs.

WEIGHT	WORD	TERM NUMBER
.7969+	CHANNEL FLOW *	2335
.4922+	INCOMPRESSIBILITY *	7281
.4298+	CONDUCTION *	2896
.3940+	INVISCID *	7627
.3308+	CHANNEL *	2336
.3085+	ONE *	10735
.3063+	STEADY *	15283
.2953+	FLUID *	5515
.2908+	COMPRESSIBILITY *	2842
.2842+	HYDROMAGNETISM *	7007
.2800+	VISCOSITY *	17561
.2622+	BOUNDARY LAYER *	1774
.2349+	LAMINAR *	8147
.2291+	MAGNETIC FIELD *	8879
.2268+	DIMENSIONAL *	3882
.2098+	GAS FLOW *	5905
.2061+	FLUID MECHANICS	5512
.2027+	MAGNETISM *	8912
.1973+	PARALLEL *	11140
.1940+	REYNOLDS NUMBER *	13434
.1883+	LAMINAR BOUNDARY LAYER	8138
.1825+	TRANSVERSE *	16798
.1777+	FLOW *	5491
.1713+	MAGNETIC *	8898
.1659+	MAGNETOHYDRODYNAMICS *	8935
.1552+	CONDUCTIVITY *	2899
.1440+	HYPERSONIC FLOW	7079
.1399+	FLAT *	5412
.1378+	TWO	17060
.1374+	CONDUCTOR *	2900
.1364+	TURBULENT FLOW	17015
.1356+	DISSIPATION *	3986
.1285+	NONEQUILIBRIUM	10392
.1274+	FLOW FIELD	5483
.1244+	TURBULENT	17019
.1225+	WAKE	17718
.1173+	CONVECTION	3026
.1138+	THREE	16421
.1100+	HYDRODYNAMICS	6964
.1032+	SUPERSONIC FLOW	15723
.1018+	LAYER	8244
.0996+	BLUNTNESS	1629
.0983+	BOUNDARY	1779
.0972+	VORTEX	17671
.0963+	WALL	17728
.0938+	FIELD	5260
.0932+	BLUNT BODY	1631
.0922+	RAREFACTION	12998
.0912+	AXISYMMETRY	1176

FIGURE 13 SECOND-GENERATION PROFILE FOR MAGNETOHYDRODYNAMIC FLOW

WEIGHT	WORD	TERM NUMBER
.0912+	SHOCK WAVE	14263
.0902+	MAGNET	8962
.0894+	NUMBER	10614
.0872+	COIL	2690
.0850+	HEAT TRANSFER	6470
.0839+	DUCT	4163
.0829+	STAGNATION	15184
.0827+	GAS DYNAMICS	5899
.0811+	GENERATOR	5968
.0778+	TURBULENCE	17011
.0763+	THERMOCONDUCTIVITY	16300
.0722+	AIRFOIL	349
.0695+	HYPERSONICS	7094
.0682+	ACCELERATOR	45
.0658+	ENERGY CONVERSION	4723
.0644+	LEADING	8254
.0641+	PLATE	11889
.0627+	INLET	7424
.0624+	SUBSONIC	15603
.0620+	TRAPPING	16808
.0614+	AIRFLOW	346
.0608+	ELECTROMAGNETISM	4503
.0603+	SHOCK	14250
.0562+	CAVITATION	2209
.0559+	MIXING	9702
.0559+	FRICTION	5745
.0548+	SHOCK TUNNEL	14255
.0529+	INDUCTION	7318
.0526+	STREAM	15451
.0523+	CURRENT	3350
.0512+	ELECTRICITY	4393
.0506+	DROP	4128
.0498+	PROFILE	12377
.0496+	HYDRAULICS	6931
.0496+	UNIFORM	17183
.0492+	ANISOTROPY	648
.0489+	DIPOLE	3913
.0482+	VELOCITY	17421
.0465+	GEOMAGNETIC FIELD	5995
.0442+	HYPERSONIC	7086
.0432+	ELECTRIC	4430
.0430+	WAVE PROPAGATION	17802
.0426+	SUPERSONIC	15730
.0425+	GAS	5924
.0422+	RADIAL	12783
.0420+	BOILING	1674
.0416+	FREE	5686
.0407+	GRADIENT	6144
.0404+	AMPLIFICATION	568

FIGURE 13 (CONTINUED)

The highest-ranked pure second-generation term is FLUID MECHANICS as shown in Figure 13.* This term is not a synonym for MAGNETOHYDRODYNAMIC FLOW. On the other hand, it probably qualifies as an analogous term. On a semantic level it seems quite correct to state that the two terms both apply to certain phenomena of fluids in motion, that conceptual overlap exists even though there are important distinctions.

Whether or not this is of any use in retrieval is not clear. There exist several thesauri whose builders thought it worthwhile to cross reference the two ideas: in the Thesaurus of ASTIA Descriptions as well as in the Bureau of Ships Technical Library Thesaurus of Descriptive Terms and Code Book, the entry for MAGNETOHYDRODYNAMICS is as follows:

MAGNETOHYDRODYNAMICS

(Motion of electrically conducting fluids in electric and magnetic fields.)

Includes:

Alfven waves
Magnetofluiddynamics
Magnetogasdynamics
Magnetohydrodynamic waves

Related Terms:

Electromagnetic waves
Fluid flow
Fluid mechanics
Magnetic pinch
Mechanical waves
Plasma physics

The STAR Guide recognizes the relationship between MAGNETOHYDRODYNAMICS and FLUID MECHANICS. But the EJC Thesaurus does not. We can only say that the "relatedness" of the ideas is supported at least in part by the recorded opinion of others: the external sources of confirmation show only that we have not deluded ourselves into perceiving an imagined relationship.

Once we accept the idea that MAGNETOHYDRODYNAMICS and FLUID MECHANICS are related, the relatedness of MAGNETOHYDRODYNAMIC FLOW to FLUID MECHANICS seems to follow more or less. We have placed ourselves in several discussions with people who think they know enough about the subject area to express an opinion. Some perceive an "analogy" relationship between MAGNETOHYDRODYNAMIC FLOW and FLUID MECHANICS, some proceed to argue the distinctions. Obtaining a reliable judgment of "near-synonymy" in non-obvious cases requires rigorously controlled experimental conditions beyond the scope of this effort. However, we feel this example illustrates the capability which the second-generation process is meant to supply and suggests some of the problems involved in developing a clear picture of its utility.

---

*MAGNETOHYDRODYNAMIC FLOW and FLUID MECHANICS co-occurred in at most five documents in the collection. The two terms occurred in 239 and 283 documents respectively.

6. A "String Synthesis" Effect. - Another effect in the second-generation list in Figure 13 is also of some interest. LAMINAR BOUNDARY LAYER, the second-ranked second-generation term -- is probably present solely because the constituent parts LAMINAR and BOUNDAR LAYER were present in the first-generation profile. The effect is "technical" and peculiar to NASA's indexing policy in that indexers of this collection actively assign constituent strings of certain multiword terms to the document when they post the document to the longer term. Because of this, as pointed out before, strong first-generation associations are found between shorter terms (e.g., BOUNDARY LAYER) embedded in longer terms (e.g., LAMINAR BOUNDARY LAYER). It is evident that the second-generation process uses these links to identify terms whose constituents are heavily weighed in the first-generation profile.

The usefulness of such a "string synthesis" effect was shown in Section C. In that example, WEIGHTLESSNESS and SIMULATION were both highly associated terms, and the desirability of finding the composite term WEIGHTLESSNESS SIMULATION was discussed.

7. Conclusions Relative to Second-Generation Associations. - The evidence is adequate to permit us to conclude that second-generation associations can contribute to the formation of more complete association profiles when there are associated terms that do not co-occur frequently enough to be detected by first-generation calculations alone, and that the idea of using second-generation associations continues to have merit. Further investigation and study could produce useful results.

The evidence is not yet adequate to provide a rational basis for using or needing the second-generation associations in practice. Nor has it been established that this avenue is the most promising one to pursue in improving the quality or the usefulness of the association data available. We consider it a valuable technique to be held in reserve for the time being.

F. 88-Term Thesaurus Development

INTRODUCTION

Recognizing that association profiles could provide aid in the current development of Thesauri for the NASA vocabulary, special purpose profiles were constructed for 88 vocabulary terms of interest. These terms are distributed over a broad frequency interval (204 to 4493) and generally range over the conceptual area of the collection.

The association profiles of the 88 terms were computed over the entire 18,000 term vocabulary, and were developed in the following manner:

a) A tape was prepared which, for each of the 88 header terms, showed its co-occurrence frequency with all other terms. For each header term, the co-occurring terms were ranked by frequency of co-occurrence, and the sequence was cut off at rank 3,000 because of computer program limits;

b) All terms which co-occurred with the header term only once or twice were deleted;

c) Association values were computed based on the square root normalization:

$$V_{ab} = fab/\sqrt{fa \cdot fb}$$

d) All associated terms with total frequencies of three or less were eliminated.

e) The threshold on association values was increased until 200 or fewer terms remained for each of the 88 header terms. Note that in some cases increasing the threshold slightly during this process caused the number of items on the list to drop well below 200.

f) Each association profile was sorted alphabetically.

Very preliminary studies suggested the use of the square root normalization for producing profiles for this particular application. It was apparent, however, that there was no theoretical foundation for this choice and that little experience was available concerning the use of various normalizations to guide the selection. Consequently, the developments described in the following sections were pursued to provide such experience and foundation.

1. Background. - Many statistical formulas, almost all of them based on the single theoretical foundation offered by the 2 x 2 contingency table, have been introduced as candidates for measuring the degree, A_{ab}, of first generation association between two index terms a and b. Generally speaking, no firm relationship between the choice of association measure and its effect upon evaluated performance has been established. In part this has occurred because no opportunity for large-scale in-use "tuning" of the measures to the specific needs of an operational retrieval system has been pursued to an experimental conclusion: The study of these measures has tended to stay in a research context, and few practice-oriented appraisals have been made. Nevertheless, a few comparative side-by-side appraisals of the effect of using different formulas have been conducted in pilot experiments.* While clear-cut preference for one formula over another (because it is a better discriminator of terms judged to be related) has not emerged from the experimental tests so far reported, the insight and experience that has been gained in laboratory tests has been valuable.

Not surprisingly, each formula has been found to have some attributes and some deficiencies. Apparently, each formula does provide, in practice, a set of associated terms among which there are many "reasonable" ones.

*See, for example, Kuhns (22) and Dennis (23).

What is annoying is that no clear-cut criterion for choice among the alternates has emerged. As a result, few candidate measures have been permanently dismissed from consideration, and a rather large set of formulas remains available.

The argument behind a typical association measure, when developed along statistical or theoretical lines offers little or no basis for distinguishing among them. The reasoning suggests comparing the number of observed co-occurrences with the calculated number of expected co-occurrences. Given two formulas, it will generally be found that substantially this same supporting rationale is proffered for both; there are many ways of measuring statistical surprise or the unexpectedness of an observation, and a large number of the available formulas can responsibly claim to do so. Figure 14 exhibits some of the more familiar measures and records their theoretical interpretation or rationale.

The fact that a large number of formulas has apparently survived the efforts of critical researchers to select among them is a curious problem which faces the serious student of associative retrieval. There definitely are differences in how the various formulas behave. But choosing which is "best", even under stated conditions, is a problem which has only rarely been approached. In this section we develop some of the tools we found helpful for comparing ranked term listings (profiles) produced by the use of term association measures, in practice.

2. Equivalent Association Measures. - In practice, the use of one of the available statistical association measures serves two purposes. The first is to select, for a given header term, a list of associated terms, a process typically accomplished by specifying a threshold for the measure above which terms count as "associated." The second application is to provide a quantitative measure of the degree of association between the associated terms and the header. A convenient way to portray the result of applying the association measure to the data is to rank the co-occurring terms in a printed list, displaying the terms in decreasing order of the association measure being used. The order in which the terms are presented on such a "profile" exhibits whether one term is more* associated with the header term than another term is.

In attempting to choose among association formulas, we postulated that our first concern was to find a measure which yields an acceptable ranking of associated terms. The magnitudes of the numerical values for the degree of association are initially of no interest. What matters is whether the most closely associated term under formula A is one of the most closely associated terms under formula B. If so, the formulas are similar; if not, dissimilar. Generalizing these ideas, two formulas which yield the same (or substantially the same) ranking of terms in the profile are equivalent from this point of view.

---

*This is true except when there are ties produced by the association measure in use.

FIGURE 14

If a and b are index terms, tallies of numbers of documents indexed (or not indexed) by the combinations of a and b are revealed in the 2 x 2 contingency table:

	a	not a	Total
b	fab	fb - fab	fb
not b	fa-fab	N-fa-fb+fab	N-fb
Total	fa	N-fa	N

where fa and fb are the frequencies of terms a and b respectively, fab is the number of co-occurrences of a and b , and N is the collection size.

Various measures based on this table are:

(I) $Aab = \frac{fab}{fb}$ — The conditional probability that given that term b is assigned to a document, term a is also assigned.

(II) $Aab = fab - \frac{fafb}{N}$ — The difference between the observed number of co-occurrences and the expected number based on chance.

(III) $Aab = Log_{10} \left(\frac{\left| fabN - fafb \right| - \frac{N}{2} \right)^2}{fafb\ (N-fa)\ (N-fb)} N$ — The chi square formula using marginal values of the 2 x 2 table and Yates correction for small samples.

(IV) $Aab = \frac{fab}{fa + fb - fab}$ — The number of co-occurrences normalized by the number of documents indexed by only one of the terms.

(V) $Aab = \frac{\left(fab - \frac{fafb}{N} \right)}{\sqrt{\frac{fafb}{N}}}$ — The number of standard deviations the observed co-occurrence falls to the right of the expected cooc.

The Derivation of Association Measures Based on the 2 x 2 Contingency Table

The notion of equivalent rankings is important principally because this is the practical way to tell the formulas apart. One prepares profiles using several formulas and examines them to see which one places the most suitable terms at or near the head of the list. Attention to the numerical values assigned is secondary. Since we are trying to relate the behavior of the formulas to the kinds of things a person comparing such profiles side-by-side would look for, the ordering is the property to examine first.

3. Generating a Spectrum of Association Measures. - The objective of comparing the ranking behavior of various formulas is well served by finding a useful way to place them all into the same mathematical form. One way to do this is to develop a general expression that generates all the formulas of interest and reduces to any specific one by a choice of parameters in the general expression. But a glance at the expressions for A_{ab} in figure 14 shows that a general expression that would include, for instance, formula III as a special case would be too complicated to manage. Fortunately, by directing attention to the approximate ranking produced by a formula, it is possible to use a simple, readily understandable model for generating a useful spectrum of alternatives.

Let term a with frequency fa be the header term for which we wish to develop a profile. Let some other term b, with frequency fb, co-occur with a fab times, as shown by the matrix in figure 15. Let us now think of a as defining (as it clearly does) a set of documents: those documents indexed by a. Let us think of the other terms b in the vocabulary (candidates for being "associated" with a) as single-term requests. And define the objective of each of these searches to be the retrieval of those documents indexed by a. In short, the a-indexed documents (and only those) are "relevant". The b indexed documents are "retrieved".

With this conceptual attitude, the familiar Recall and Precision measures can be defined for each term b (with respect to the given term a). They measure - with the usual disclaimers - the goodness of b as a substitute for a.

The recall of term b is the proportion of documents indexed by term a which are also indexed by term b:

$$\text{Recall}_b = \frac{fab}{fa} \qquad (1)$$

The precision of term b is the proportion of documents posted to b which are also indexed by term a:

$$\text{Precision}_b = \frac{fab}{fb} \qquad (2)$$

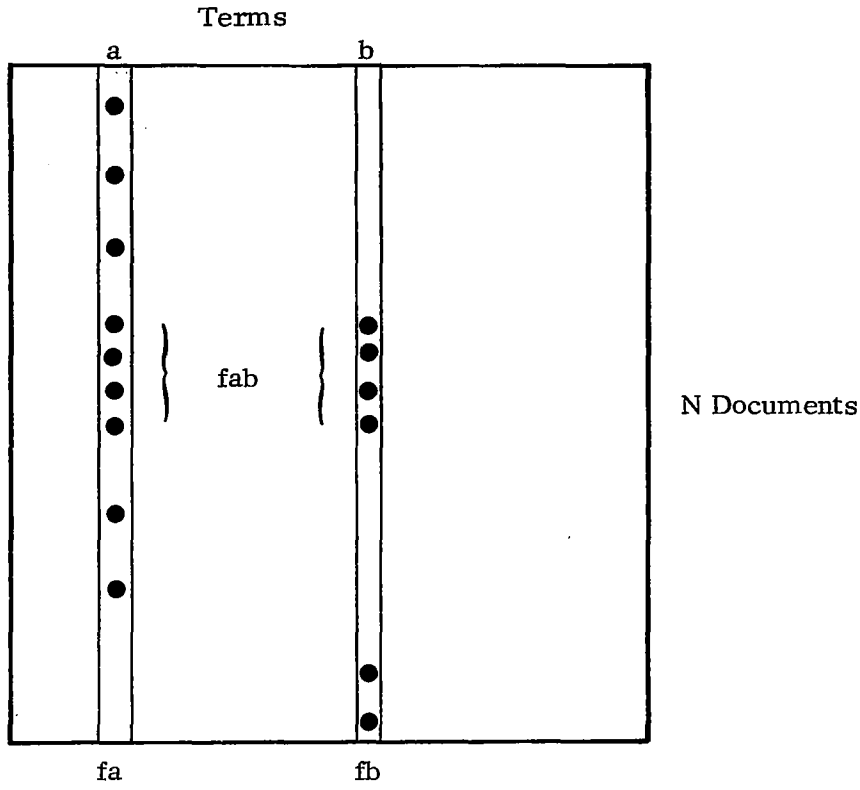


FIGURE 15 DOCUMENT-TERM MATRIX SHOWING THAT TERM A (FREQUENCY fa) CO-OCCURS WITH TERM B (FREQUENCY fb) EXACTLY fab TIMES

We now have two measures of b's capability to be used in lieu of a; since the degree to which b is associated with a is a single number, we wish to combine Recall and Precision into a single measure. The product suggests itself since a term b with both high Recall and Precision should have a high value. But since we have no idea whether to consider Recall more important than Precision or vice versa, we multiply them together with adjustable exponents. Thus a spectrum of directly interpretable measures of the association of term b with term a is provided by:

$$Aab = \left(\frac{fab}{fa}\right)^{(1-n)} \left(\frac{fab}{fb}\right)^{n} \qquad (3)$$

where n is such that $0 \leq n \leq 1$. Varying n generates a variety of association measures and represents a different interpretation of the relative importance of Recall and Precision in this viewpoint towards associations.

Since we ascribe little merit to the actual numerical value Aab given by a measure, putting emphasis rather on the resulting profile term ranking, we allow ourselves to alter a given measure including this one so long as the ranking remains invariant under the alteration. Two alterations of this kind which yield equivalent rankings are important:

a) Any positive power of a formula which yields nonnegative association measures produces the same ranking of terms as the original formula:

Proof: $A_{xy} > A_{xz} > 0$ and $k > 0$ implies $A_{xy}^{k} > A_{xz}^{k} > 0$

b) Also, we will usually be allowed to strike the factor fa from the measure, since it is the same constant for all the terms in a's profile and therefore does not affect the ranking.

Applying these rules to (3) yields the equivalent formulation:

$$A_{ab} \approx (fab)^{(1-n)} \left(\frac{fab}{fb}\right)^{n} = \frac{fab}{fb^{n}} \qquad (4)$$

The model presented above thus yields a spectrum of measures of the type $Aab = \frac{fab}{fb^{n}}$. Each such measure is "rank-equivalent to" (produces the same ranking as) a formula derived from a specified weighting of Recall and Precision in this framework.

The next task is to find, for the more complex statistical formulas, which choice of n produces substantially the same ranking of terms. This will allow us, if we choose, to interpret those other formulas in a common framework. Let us therefore examine more closely the way the choice of n affects the ranking and develop some of the apparatus for relating n to more complex measures.

Figure 16 shows a graph of the (fb, fab) space which is of interest because of the form of (4). Each term which co-occurs with the header term can be placed as a point on this graph according to its frequency (fb) and the number of times it co-occurs with the header term (fab). (Note that all terms must be located on or below 45° line, since fb $\geq$ fab.) An association measure is a curve of stated shape which moves in this space, and the ranking of terms on a profile according to that measure is the order in which this curve passes points which represent co-occurring terms.

Viewing the measures in this way allows us to perceive which areas of the space are passed first and, therefore, which terms are likely to be highly ranked.

In figure 17 we have shown the curves and movements for various values of n in Equation (4).

When n = 1, we have a straight line rotating clockwise about the origin. This is the representation of the measure

$$Aab = \frac{fab}{fb}$$

which is pure Precision in terms of our model. Note that the maximum value attainable arises when fab = fb, i.e. term b co-occurs with term a each time it occurs. Thus, a term with the values fb = 1, fab = 1 or equivalent must be ranked #1. No distinction is made between such a term and one with values fb = 10, fab = 10, although there seems to be more evidence of association in this latter case.

The graphical representation of the case for n = $\frac{1}{2}$ is a curve as shown in figure 4 again rotating clockwise about the origin. The measure for n = $\frac{1}{2}$ is

$$Aab = \frac{fab}{(fb)^{\frac{1}{2}}}$$

and may be contrasted to the case for n = 1. Now, of course, a term with values fb = 10, fab = 10 would be ranked higher than a term with values fb = 1, fab = 1. This phenomenon is seen by inspection of the areas covered first by the movement of the corresponding curves. The bend in the curve for n = $\frac{1}{2}$ causes it to be well above the point fb = 1, fab = 1 when it crosses the point fb = 10, fab = 10. In general, measures of the type fab divided by some root of fb tend to bend more sharply as n approaches 0.

The limiting case when n does reach 0 is given by a horizontal line straight downward, which moves decreasing fab. This measure ranks the terms in order of co-occurrence count. (This is pure Recall.) Not all possible ranking rules fall within the scope of the model, of course.

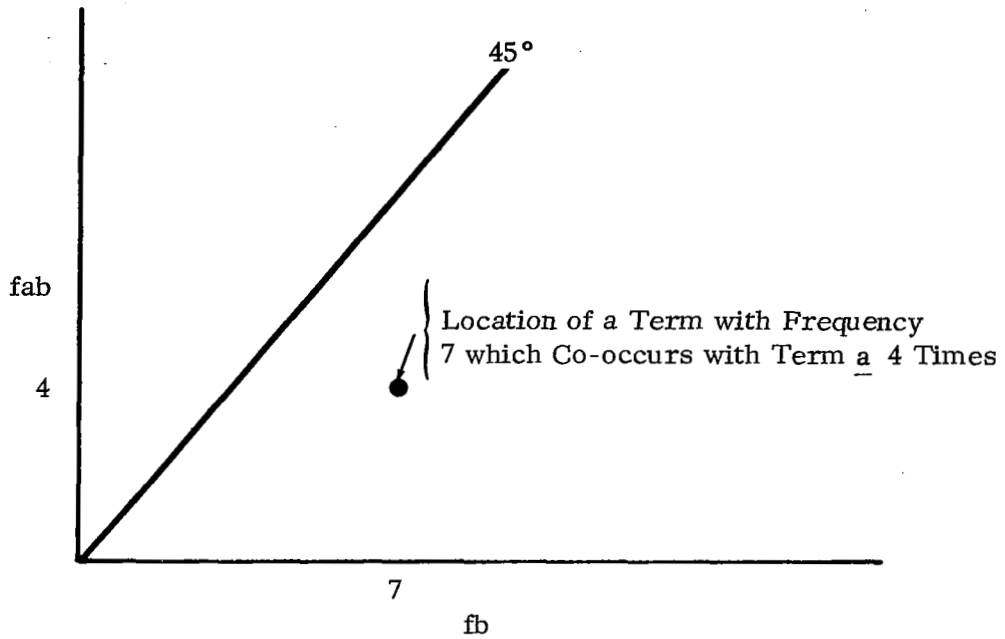


FIGURE 16 THE (fa, fab) SPACE

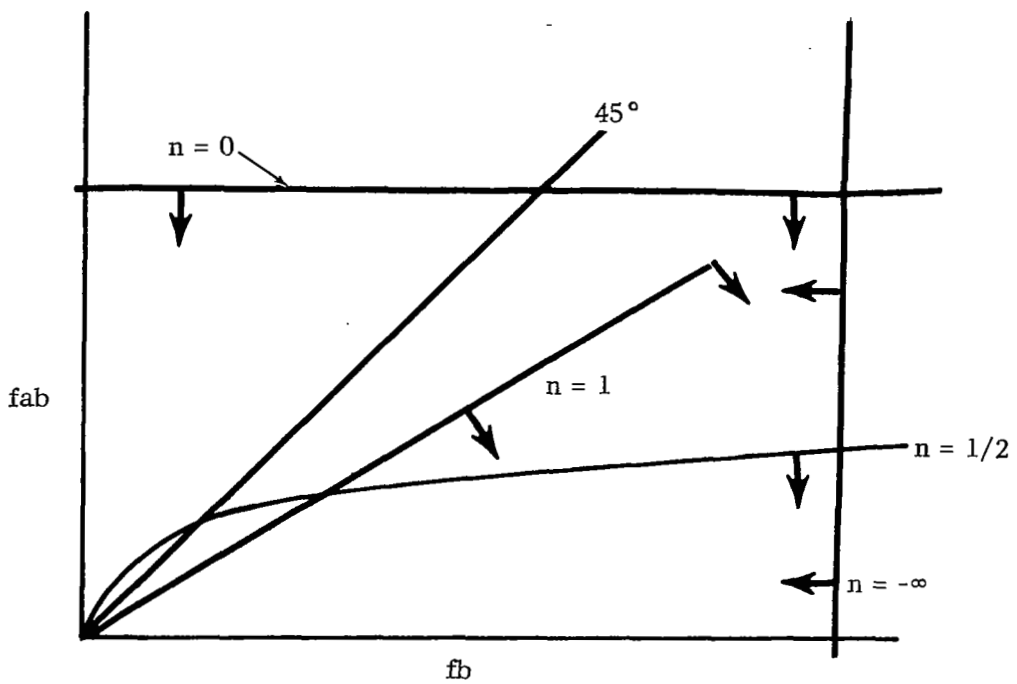


FIGURE 17 GRAPHICAL REPRESENTATION OF VARIOUS MEASURES SUGGESTED BY THE MODEL

For example, the extreme case represented by a vertical line that moves from right to left is not strictly within the scope of the model, since n would have to be $-\infty$ It is of interest as an extreme, however, and for this reason is shown in figure 17. It corresponds to the measure

$$Aab = fb.$$

To the extent that it is important and useful to identify areas in the space which are preferred by different values of n, it would be useful to know the exact distribution of the points in that space. One could then not only identify areas of preference, but what density of points in that area is.

For example, one measure might sweep across a particular area in contrast to another area for a different measure. However, if the two areas in question are extremely sparse, the measures might yield very similar rankings. In fact, two quite different measures will yield the same ranking if the terms in the (fb, fab) space are arranged in some particular fashion. Thus, knowledge about the distributions of fb and fab is valuable.

Figure 18 indicates what the distributions of fab and fb might be for the terms which co-occur with a given header term. The combination of these distributions yields a probability density function, representable by the shaded portions of figure 18. It would be possible empirically to derive the distribution of (fb, fab) by sampling a large portion of the data, and thus obtain the probability density function previously mentioned.

It is known by observation that the points are distributed in the manner shown in figure 18, with a very high density of points in the lower left hand corner, trailing off horizontally and upwards. For the present purposes we need know little more.

4. Relationship With Other Measures. - In general the shape of a curve corresponding to some given association formula that is a function of (fab, fb) can be obtained by setting the measure equal to a constant. This constant represents a particular threshold, and the movement of the curve corresponds to alteration of the threshold. As the threshold is reduced, the curve moves downward and a ranking results.

The measure given by

$$Aab = fab - \frac{fafb}{N} \qquad (5)$$

where N = collection size, has been suggested by Maron and Kuhns (5). When set to a constant (to represent a particular threshold K), and solved for fab, (5) becomes

$$fab = K + \frac{fafb}{N}$$

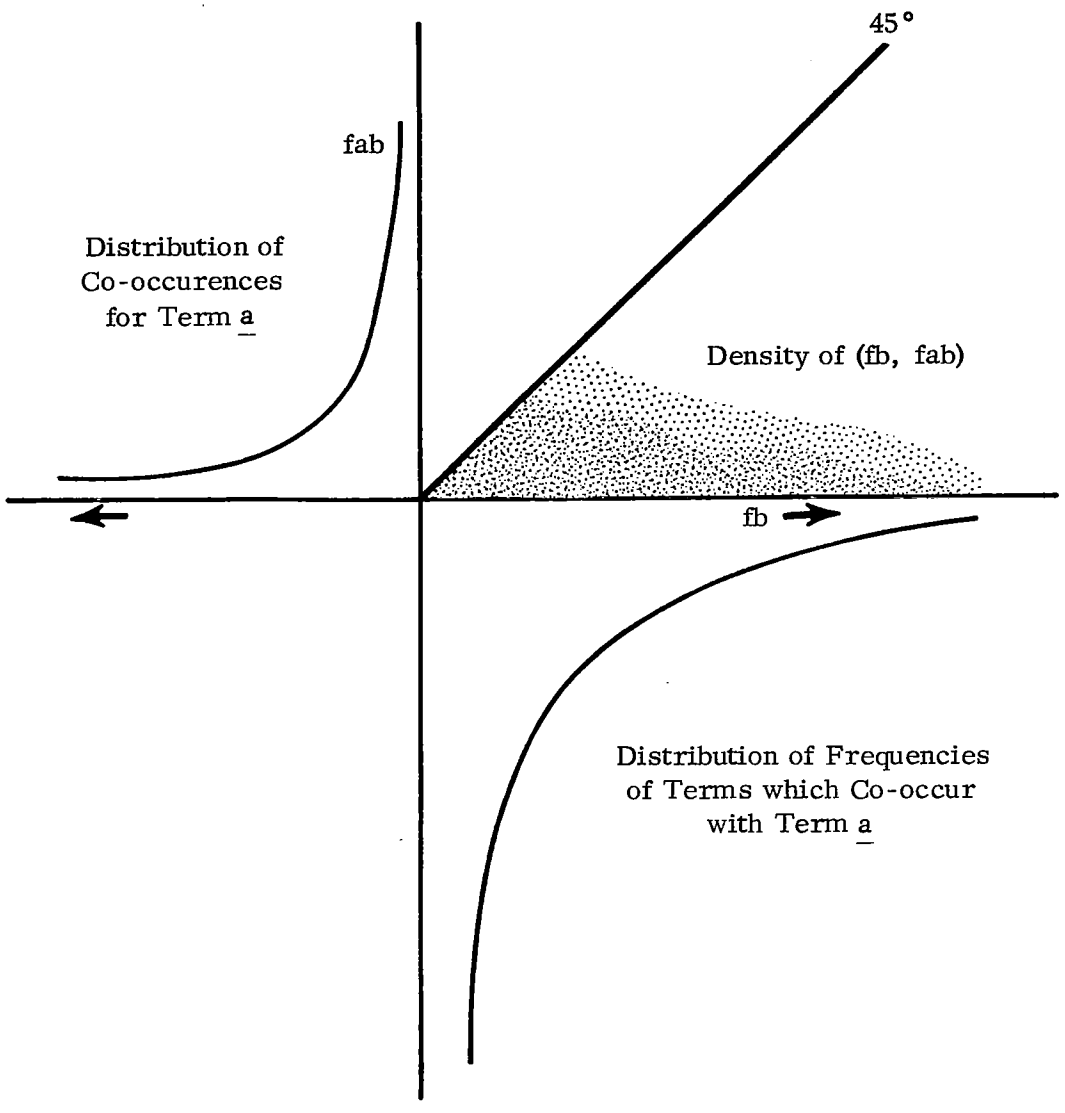


FIGURE 18 POSSIBLE DISTRIBUTION OF POINTS IN THE (fb, fab) SPACE

The graphical representation of fab as a function of fb is thus a straight line with small* slope fa/N, emerging from the vertical axis at the point fab = K.

As the threshold is reduced, this nearly horizontal* line moves vertically downward. The ranking it produces can be expected to be very similar to a ranking by fab alone, except that the slope of the line is fa/N rather than zero. Note that we cannot strike fa or N from the formula as they are not pure additive or multiplicative constants.

A similar graphical interpretation is given for

$$Aab = \frac{fab}{fa + fb - fab}, \qquad (7)$$

a measure suggested by Doyle (16).

When set to K and solved for fab we get

$$fab = \frac{K}{K + 1} fb + \frac{K}{K + 1} fa \qquad (8)$$

Again a linear relationship exists between fb and fab as in the previous case. However, the slope of the line approaches zero as the line moves from the top of the graph (where the slope is 1) to the bottom.

In the formula suggested by Dennis (23) and given by

$$Aab = \frac{fab - \frac{fafb}{N}}{\sqrt{\frac{fafb}{N}}} \qquad (9)$$

we may eliminate fa and $\frac{1}{N}$ as constants. Setting Aab to K and solving for fab yields

$$fab = K\sqrt{fb} + \frac{fa}{N} fb \qquad (10)$$

this is seen to be representable as the sum of the square root curve (i.e., n = ½) and a straight line with slope fa/N. Since we would expect fa/N to be small, (9) should yield a ranking similar to the case when n = ½. A measure investigated by Stiles (3) and given by

$$Aab = \log_{10} \frac{\left(|fabN - fafb| - \frac{N}{2}\right)^2 N}{fafb\ (N-fa)\ (N-fb)} \qquad (11)$$

*Only a handful of terms a (in the NASA vocabulary, 10 of 18,000) have fa/N in excess of .03.

is approximately equivalent to the ranking produced by the measure

$$\frac{(fab - \frac{1}{2})^2 N}{fa\,fb} \qquad (12)$$

when fa x fb is less than N*. Striking out N and fa we can see that this measure will approximate the case when $n = \frac{1}{2}$. The representation of this measure then will be a curve very similar to

$$fab = K - fb \qquad (13)$$

The formulas treated above (except for Doyle's) are thus convertible, without extraordinary difficulty into a form where a value of n in $\frac{fab}{fb^n}$ can be assigned as a crude descriptive parameter. We do this because of a desire to compare the formulas within the simpler framework provided by the Recall and Precision model discussed earlier. The relations will be clearer after studying the illustrative examples in the next section.

5. Illustrative Example. - The relationships among the various formulas discussed in the preceding section are illustrated in the example presented in Figure 19. For reasons of space and clarity, the length of the association lists is radically curtailed. The data are drawn from the NASA collection statistics, a collection we noted earlier which contains about 100,000 documents and 18,000 index terms.

Figure 19 illustrates the top 15 associates for the header term "Rocket Motor Case" as ranked by each of 9 measures. The set of co-occurring terms was first restricted to include only those terms b such that 2% or more of their occurrences were co-occurrences with "Rocket Motor Case", i.e., $fab/fb \geq .02$. This restriction cut the set of candidates to about 1/3 of the set of all terms that co-occurred with the header term, and was necessary because of computer program limitations. This selection governs all the lists compared in figure 19.

Five of the rankings shown are produced by arranging the term b according to five choices of n in equation 4. The remaining 4 rankings were produced by measures in figure 14.

Figure 20 shows the position of the curve representing each of the measures as it selects its 15th term. That is, in the positions indicated, each of the measures has chosen 15 terms associated with "Rocket Motor Case". The thresholds corresponding to these positions vary, and for a particular measure the threshold is merely the value of that measure for the term ranked 15. Thus, by setting the measure equal to the threshold represented by the 15th ranked term, the equation of the curve at that point results.

---

*See Fossum (13)

FIGURE 19

The 15 Top Associates of the Term ROCKET MOTOR CASE
As Ranked by Each of Nine Association Measures

fb ($n = -\infty$)	fab (n=0)	$fab - \frac{fafb}{N}$
Rocket	Case	Case
Propellant	Motor	Motor
Solid	Rocket	Rocket
Steel	Rocket Engine	Rocket Engine
Rocket Engine	Solid	Solid
Fabrication	Propellant	Propellant
Titanium	Steel	Steel
Motor	Fabrication	Fabrication
Glass	Winding	Winding
Grain	Filament	Filament
Insulation	Solid Prop. Rocket Eng	Solid Prop. Rocket Eng
Welding	Filament Winding	Filament Winding
Fracture	Titanium	Titanium
Bonding	Fiber	Fiber
Cryogenics	Glass	Glass

STILES	$fab - \frac{fafb}{N}$	$\sqrt{\frac{fafb}{N}}$	$\frac{fab}{(fb)^{1/2}}$ $(n = 1/2)$
Case	Case	Case	Case
Motor	Motor	Winding	Motor
Winding	Winding	Rocket	Winding
Rocket	Filament Winding	Filament Winding	Rocket
Filament Winding	Deep Draw	Stretch Forming	Filament Winding
Stretch Forming	Rocket	Deep Draw	Deep Draw
Deep Draw	Stretch Forming	Hydrotest	Stretch Forming
Hydrotest	Hydrotest	Filament	Hydrotest
Filament	Filament	Rocket Engine	Filament
Rocket Engine	Rocket Engine	Closure	Rocket Engine
Closure	Closure	Fiberglass	Closure
Fiberglass	Fiberglass	Stretch	Fiberglass
Stretch	Stretch	Steel	Stretch
Steel	Spiral Wrap	Fabrication	Spiral Wrap
Fabrication	Steel		Steel

FIGURE 19 (continued)

$\frac{fab}{fa+fb-fab}$	$\frac{fab}{(fb)^{2/3}}$ (n=2/3)	$\frac{fab}{fb}$ (n=1)
Case	Case	Altair Missile
Motor	Deep Draw	Polaris A2A Missile
Winding	Motor	Seepage
Filament Winding	Spiral Wrap	Spiral Wrap
Filament	Stretch Forming	Stretch Project
Closure	Hydrotest	TU 290 Motor
Fiberglass	Seepage	Turks Head Mill
Glass Fiber	Winding	Deep Draw
Stretch Forming	Filament Winding	Environmental Temp.
Stretch	Altair Missile	Fuzz
Solid Prop. Rocket Eng.	Stretch Project	Helical Winding
Rocket Engine	Closure	Stretch Forming
Reinforced Plastic	TU 290 Motor	Hydrotest
High Strength	Turks Head Mill	Vasco Jet
Fabrication	Polaris A2A Missile	Wing IV Motor

The various term lists shown in figure 19 are arranged systematically according to the average slopes of the corresponding curve in figure 20, beginning with the vertical line (fb) and rotating counter-clockwise until we reach the 45° line. This arrangement corresponds to increasing n from $-\infty$ to + 1 for those measures derived from the model; the other measures are interspersed.

Inspection of the lists in figure 19 will quickly indicate that they all contain a good proportion of terms which most evaluators would judge to be associated with the header term, Rocket Motor Case. This is not particularly surprising in a vocabulary of 18,000 terms. There are probably 150 good associates for each middle frequency term in a vocabulary of this size. Even if it were practical to print the top (say) 200 terms here, side-by-side appraisal of the comparative rankings would require more effort than the reader would expect (or want to devote). Fortunately, however, we can characterize each list by describing the types of terms which tend to appear at the top of the list. These characteristics are features of the measure which generated the particular list. Given a specific application for which the list is to be used, we could then hope to assess in advance which measures would be expected best to meet the application's requirements. Essentially, all that is needed is some statement about which characteristics of highly associated terms are desirable for the given application.

The list for $n = -\infty$ i.e., ranking by fb alone, is surprisingly good, even when we recall that 2% selection restriction mentioned above. That is, merely selecting the high frequency terms which co-occur with the header more than 2% of the time -- then ranking them by decreasing frequency -- yields a list of words that is far from ridiculous. This indicates, in fact, that the term co-occurrence phenomenon is a stronger effect than one might be predisposed to suspect. Naturally, this list contains terms which tend to be very general, broad, highly used vocabulary terms (by construction). It has the noteworthy attribute that there are hardly any terms appearing on this list which one needs special knowledge to understand. (The vertical line representing this measure was at fb = 637 when the 15th term was chosen.)

The terms on the list for n = 0, i.e., ranking by fab, are quite similar to those on the list for $n = -\infty$. Note, however, that the constituent terms "case" and "motor" have moved into prominence on this list. (The horizontal line which represents this measure had the equation fab = 31 when the 15th term was chosen.)

The ranking of the top 15 terms for the measure fab - $\frac{fafb}{N}$ is precisely that of the previous measure, fab. Thus the factor $\frac{fafb}{N}$ was not great enough to influence the ranking up to this point. Note the line at this point has already turned clockwise to a very small slope, the equation of the line being

$$fab = 28 + .00243fb.$$

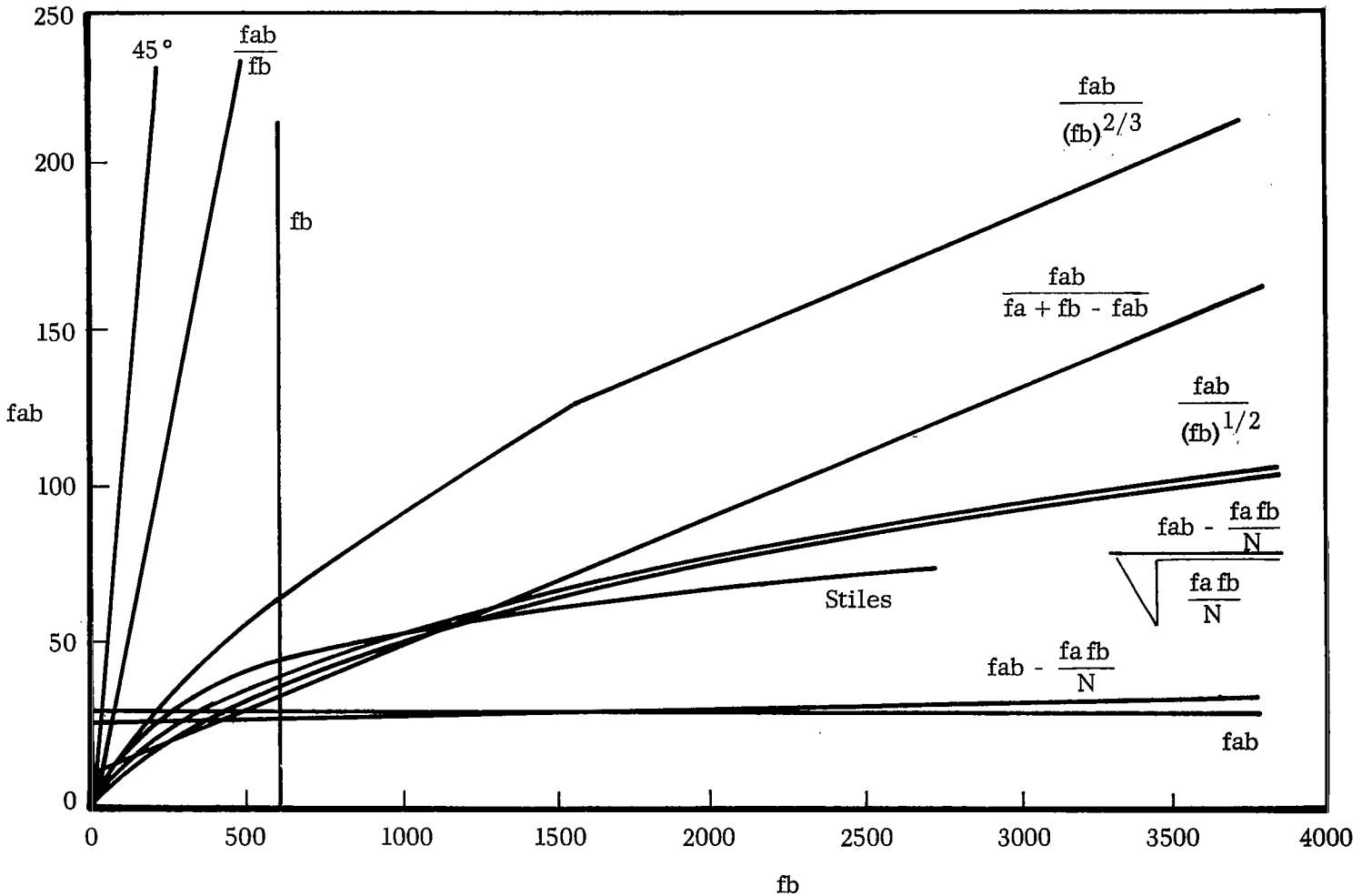


FIGURE 20 GRAPHICAL REPRESENTATION OF VARIOUS ASSOCIATION MEASURES AS THEY SELECT TERM NO. 15

The list given by the measure suggested by Stiles differs significantly from the previous lists. Technical terms, like "Hydro test", "deep draw" and "closure" begin to be included. This list and the next two are extremely similar, with only minor permutations of terms. (The curve for this measure was obtained by plotting actual points since the equation is quite complex.) However, figure 20 exhibits the obvious similarity of this and the next two measures, showing that the approximation in equation 12 is indeed valid for this header term. The equations of the curves for the measures given by

$$\frac{fab - \frac{fafb}{N}}{\sqrt{\frac{fafb}{N}}} \quad \text{and} \quad \frac{fab}{\sqrt{fb}} \quad \text{are}$$

$fab = (1.58)\sqrt{fb} + .00243fb$ and $\sqrt{fab} = (1.73)\ fb$ respectively

The ranking by $\frac{fab}{fa + fb - fab}$ contains many of the previously seen terms, plus some specific additions such as "high strength" and "reinforced plastic". The curve representing this measure is given at this point by

$$fab = .041fb + 10.$$

Next, the list for n = 2/3, i.e., $\frac{fab}{fb^{2/3}}$, shows the addition of some very specific, highly technical terms such as "seepage", "Polaris A2A missile" and "Turks head mill". The equation of the curve when it has chosen the 15th term is

$$fab = .9fb^{2/3}.$$

Finally, the list given by $\frac{fab}{fb}$ when n = 1 is presented. Almost all 15 terms on the list are low frequency, highly specific terms. The equation of the line is

$$fab = .46\ fb$$

at this point.

The discussion above, in conjunction with an inspection of the curves representing various measures, shows that the lists corresponding to various n tend to become more specific and technical in nature as n goes from $-\infty$ to + 1.

Comments on the usefulness of the 88 term thesaurus by NASA personnel are expected to provide additional data relating to a choice of normalizations for profile preparation.

PART III

THE NASADL SYSTEM

A number of programs have been written to carry out associative retrieval using large files as data bases. The overall approach has been to reduce the data base files to a size and format adaptable for processing. An associative relationship among the keywords of the file is then established and retrieval is attempted on the associated keywords and, subsequently, on the reduced data base itself.

The NASADL programs have all been written either in FAP for the IBM 7090 or in Autocoder for the IBM 1401 with a minimum of 8K of core. Listing and retrieval are carried out on the latter machine while all the preliminary file handling is done on the former machine.

The NASADL system can be divided into four fundamental operations:

1. File generation
2. File update
3. Associative matrix generation
4. Retrieval

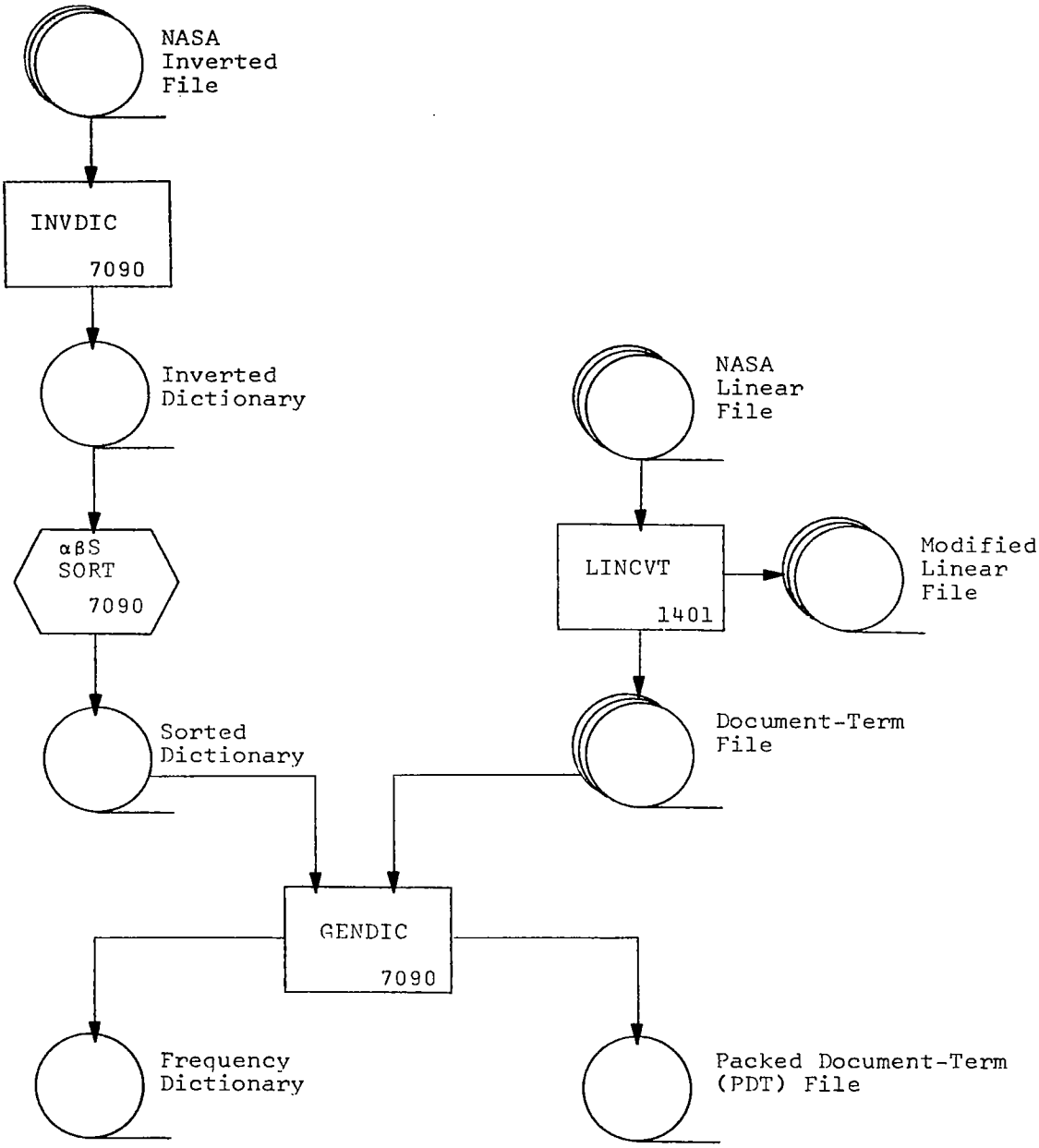
The first three operations each involve a number of program runs. The fourth operation is the running of the 1401 retrieval program.

A. File Handling and Matrix Generation

1. File Generation - The primary input data to the NASADL system is the Linear File and the Inverted File provided by NASA. In this sequence of runs (see Figures 21-23), this data is converted to a format suitable for further process and two main files are generated: a dictionary file with the frequency of occurrence of each keyword in the corpus of document abstracts and a term-term frequency file with the frequency of co-occurrence of every pair of keywords in the corpus. Additionally, the Linear File is modified for later printout by the retrieval program.

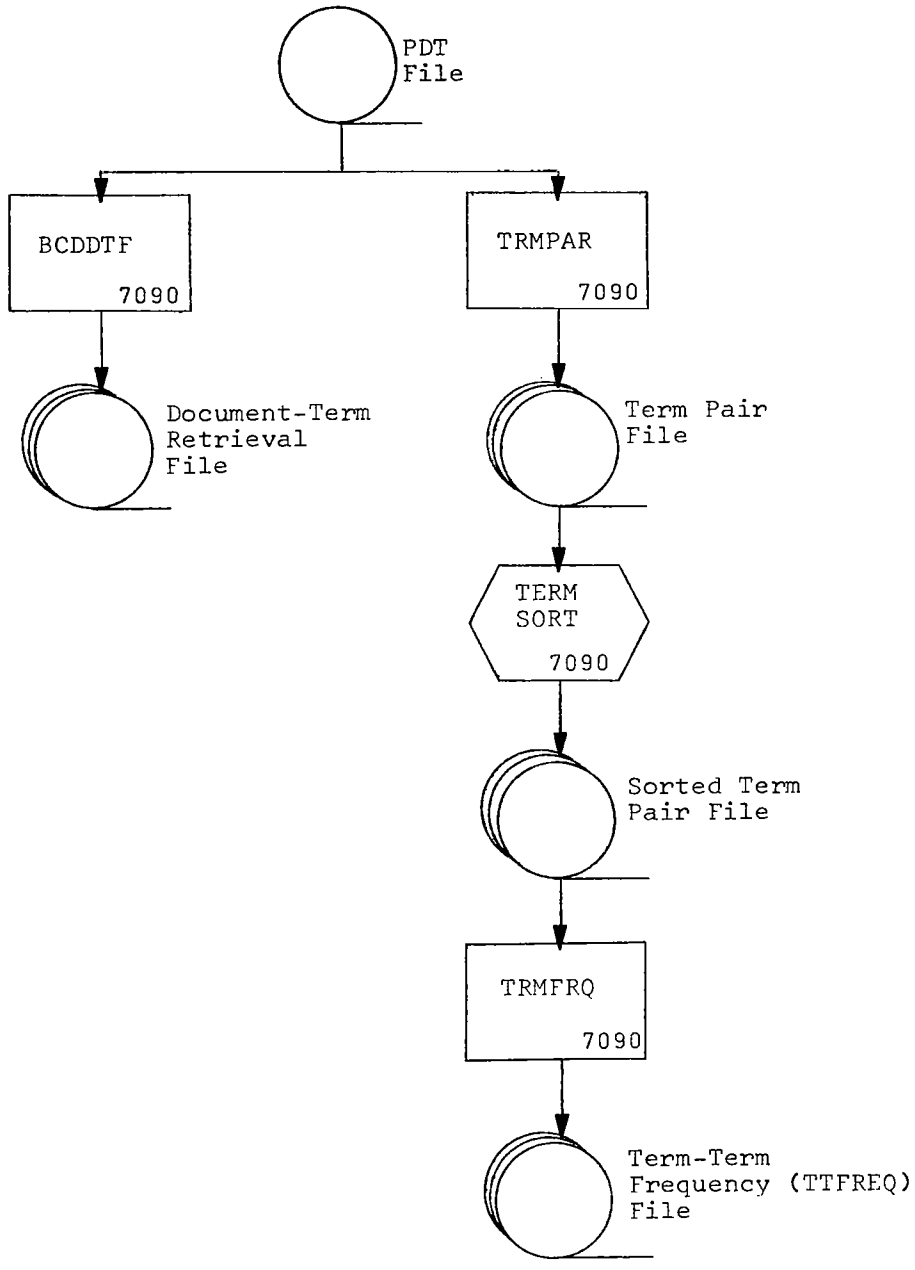
The Inverted File is processed by INVDIC. This program selects every unique keyword and writes it out on a separate file, creating a source dictionary. This dictionary is then sorted into alphabetic scientific ($\alpha\beta S$) order for processing by the 7090.

The second input to the main 7090 run is prepared by the 1401 program LINCVT. This program processes the NASA linear file and produces two outputs. The first is a formatted linear file which is later read by the retrieval program. The second is a file of document terms with all the keywords of each document included.



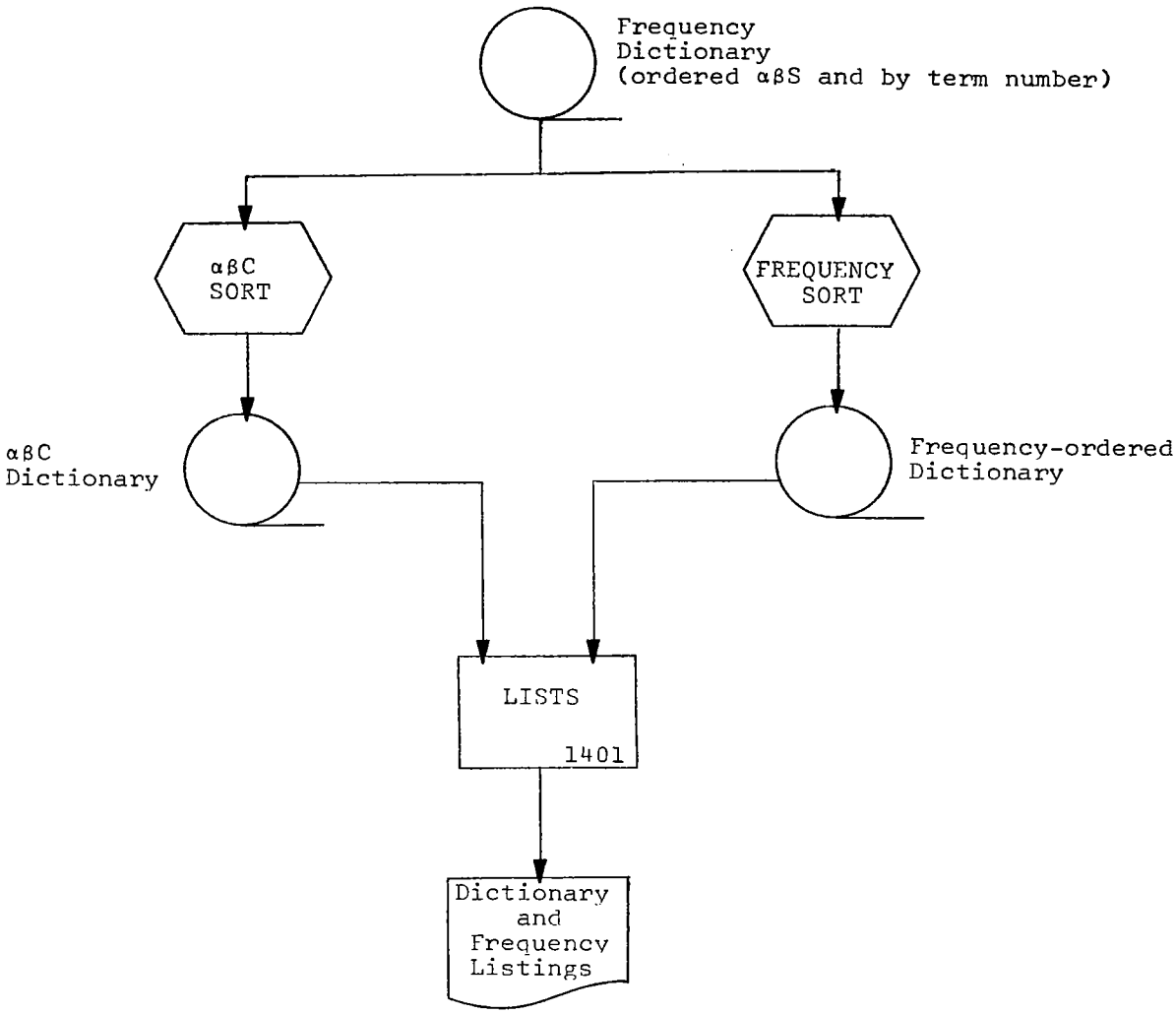
File Generation (Part 1 of 3)

Figure 21



File Generation (Part 2 of 3)

Figure 22



File Generation (Part 3 of 3)

Figure 23

GENDIC, using the two inputs discussed above, prepares a new dictionary file with the frequency of occurrence of every keyword appended to the keyword. It also prepares a file of all documents with the term numbers rather than the terms themselves appended. This process to this point reduces a six-tape Linear File to approximately a one-half tape packed document-term (PDT) file.

The PDT file produced by GENDIC is used to prepare a BCD file of the same general format which can be processed by the retrieval program. It is also used to create by TRMPAR a large file of pairs of keywords co-occurring in the documents. This file is then sorted into ascending sequence and reduced by TRMFRQ to a file of co-occurring keywords, the term-term frequency (TTFRFQ) file. This file is the original master file for the NASADL system.

The frequency dictionary prepared by GENDIC is in alphabetic order (scientific sequence) and also, by definition, in term number order. This file must be sorted into commercial sequence for the 1401 retrieval program. It can also be sorted by frequency of occurrence.

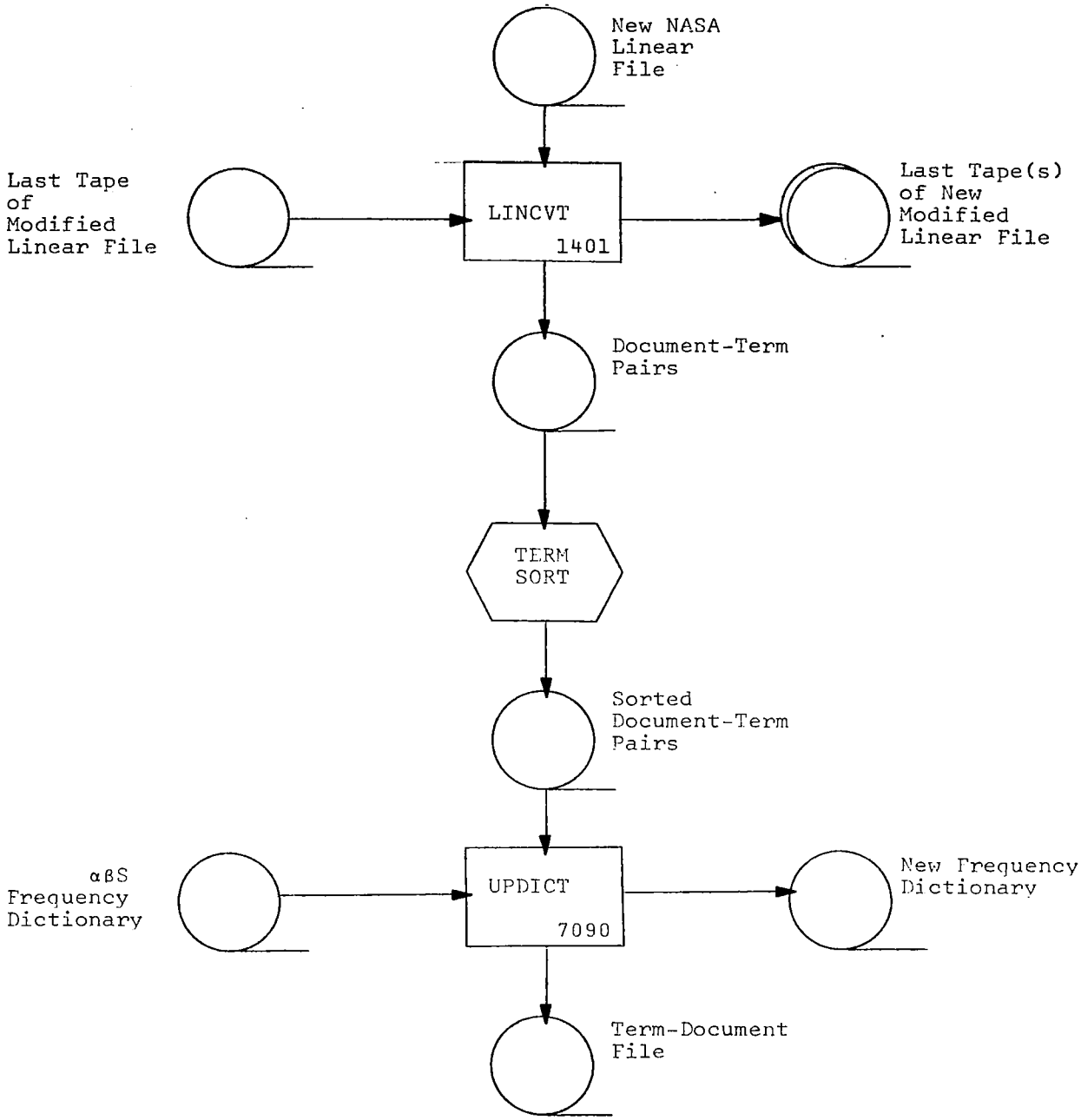
A 1401 program called LISTS can list either of these two dictionaries in a double-columnar format. LISTS can also print a summary by frequency of occurrence of the set of keywords from the frequency-ordered dictionary.

2. File Update - Once a master dictionary file and TTFREQ file have been created, a separate update procedure must be followed to add new documents to the NASADL system. (See Figures 24-27.) Two files prepared in the generation phase, the modified Linear File and the document-term retrieval file, are updated. The two other files, the dictionary and the TRFREQ, are completely rewritten.

A second option in the LINCVT program takes the new documents from additions to the NASA Linear File, reformats them, and copies them onto the end of the last tape of the modified linear file. At the same time, LINCVT also prepares a file of all document-term pairs which is later sorted into scientific alphabetic sequence by term.

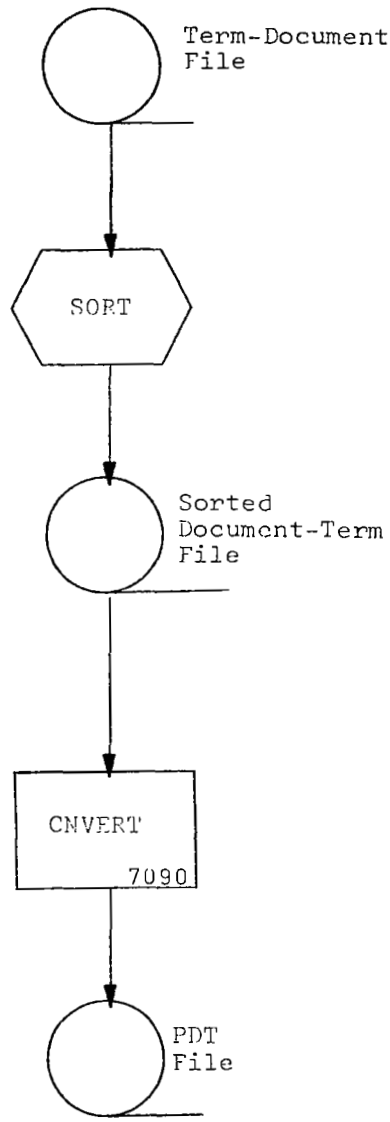
UPDICT, a parallel to GENDIC, reads this sorted file and the old frequency dictionary (in scientific sequence) and updates the dictionary frequencies. It also adds the new terms to the end of the dictionary and assigns new term numbers. A second output of UPDICT is a term number-document file.

This latter file is ordered by a sort routine by document and within document by term number. This file is then converted into the PDT format to allow for processing that parallels the file generation run.



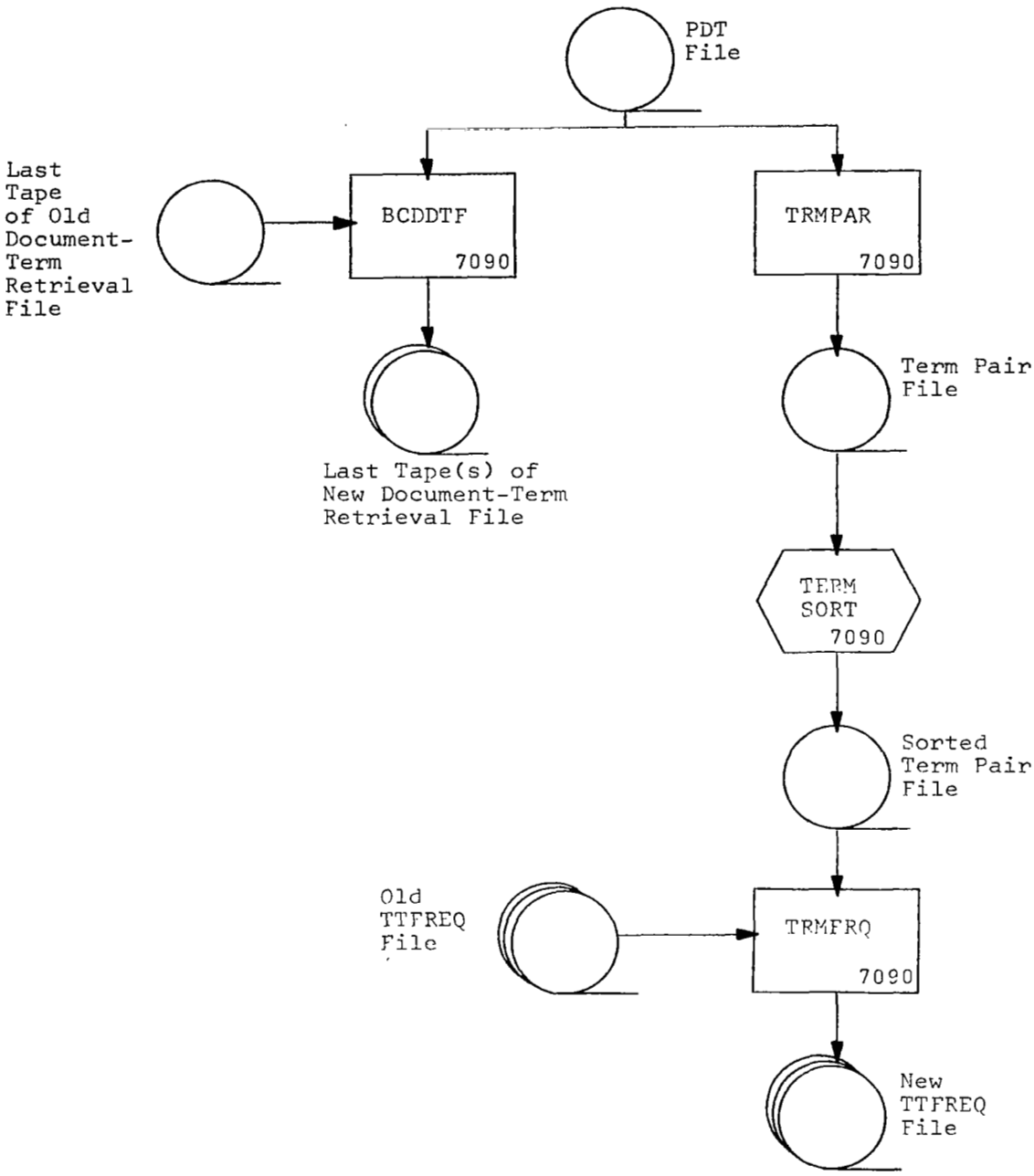
File Update (Part 1 of 4)

Figure 24



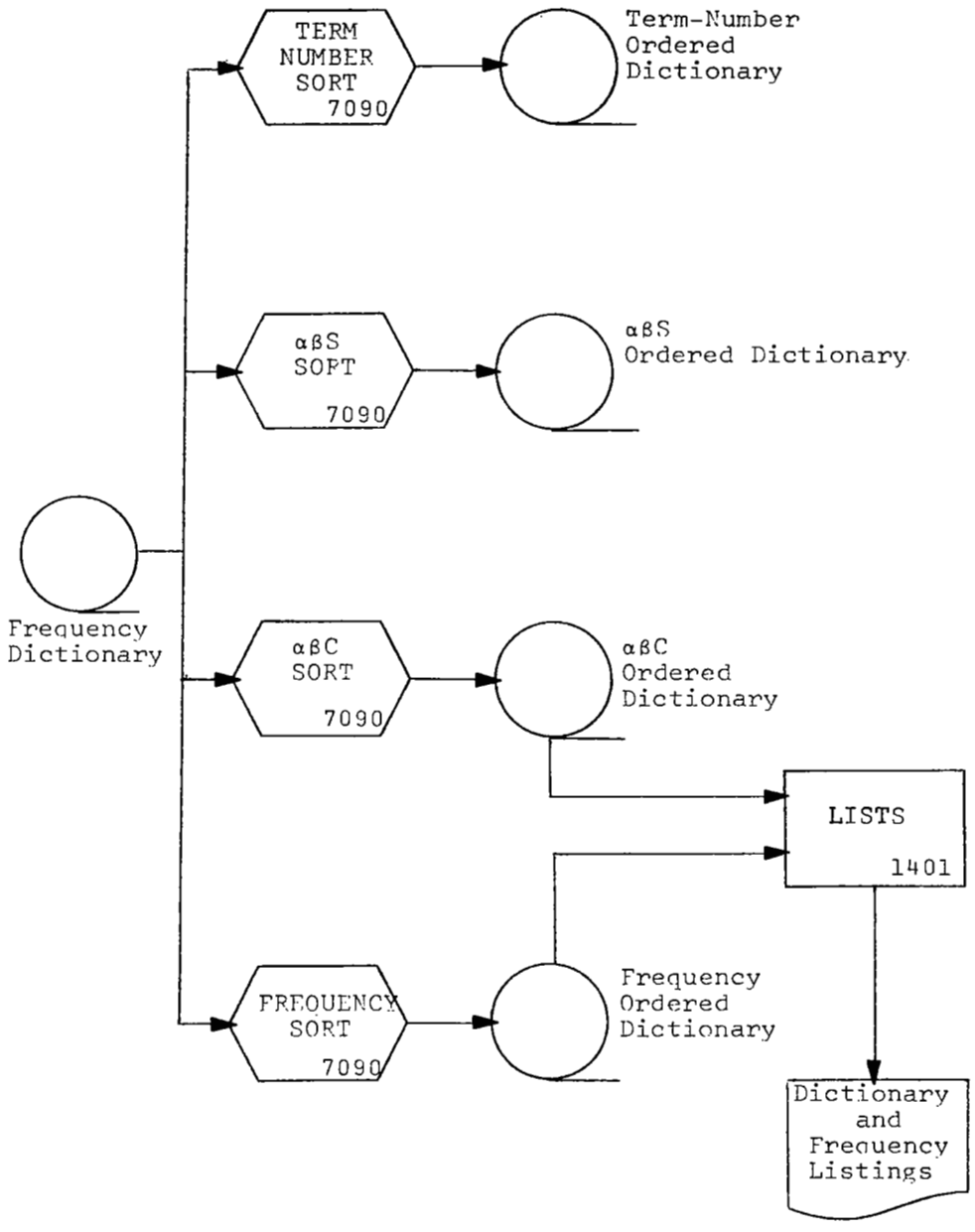
File Update (Part 2 of 4)

Figure 25



File Update (Part 3 of 4)

Figure 26



File Update (Part 4 of 4)

Figure 27

Figure 26 is quite similar to Figure 22. The operations differ only in that two of the programs, BCDDTF and TRMFRQ, are run with an update option rather than a generation option. The BCDDTF update process is trivial. The new document numbers with their term numbers are merely appended to the end of the data on the last tape of the existing file. TRMFRQ must update in a more comprehensive manner, adding additional co-occurrences of existing term pairs, adding initial co-occurrences for term pairs where the individual terms are already in the system, and lastly adding co-occurrences for terms where one or both have just been added to the system.

The new dictionary produced by UPDICT is not ordered, and four sorts are provided to produce the four desired sequences: term-number ordered and alphabetic commercial sequence for retrieval, alphabetic scientific sequence for future updating, and frequency-ordered sequence for printing.

3. Associative Matrix Generation - A general associative matrix is of the form $\lambda^{1/2}(I-\tilde{K})^{-1}\lambda^{1/2}$ where the $\lambda^{1/2}$ are pre- and post-normalizing functions. The inversion of such a matrix where the matrix dimensions are of the order of 1000 or more is not practical, and thus it is approximated by the equality

$$(I-K)^{-1} = I + \tilde{K} + \tilde{K}^2 + \tilde{K}^3 + \ldots \quad .$$

Since matrix manipulation of any form is an expensive operation, the submatrix K is extracted from the TRFREQ file by considering only those terms that have individually occurred more than some given threshold of times. The generation process is shown in Figures 28 - 30.

The first part of the matrix generation process is to select the subset of terms (a maximum suggested number is 1000) that will be used. This is accomplished by DNSMAP which also reassigns consecutive matrix row numbers to the terms kept. It must also provide a map so that the original term numbers can be later restored. This dense matrix is then multiplied by the normalizing matrix with the program NORMAL.

At this point, some approximation to the inverse of this matrix is obtained. Figure 29 shows the technique for arriving at the approximation

$$I + \tilde{K} + \tilde{K}^2 + \tilde{K}^3 = W \simeq (I-\tilde{K})^{-1}.$$

In Figure 31 are depicted the suggested processes for arriving at the approximations $W = I + \tilde{K}$ and $W = I + \tilde{K} + \tilde{K}^2$. By a similar approach, the series can be extended out as far as the computational purse string will allow.

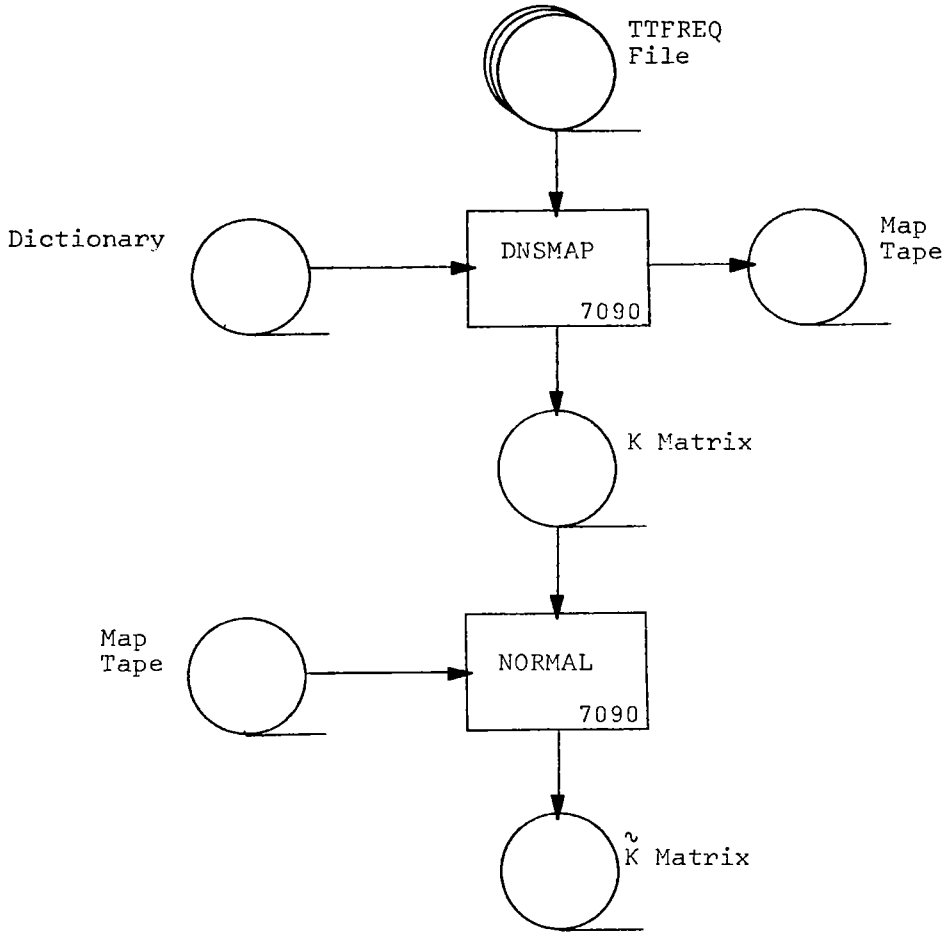


FIGURE 28 ASSOCIATIVE MATRIX GENERATION (PART 1 of 3)

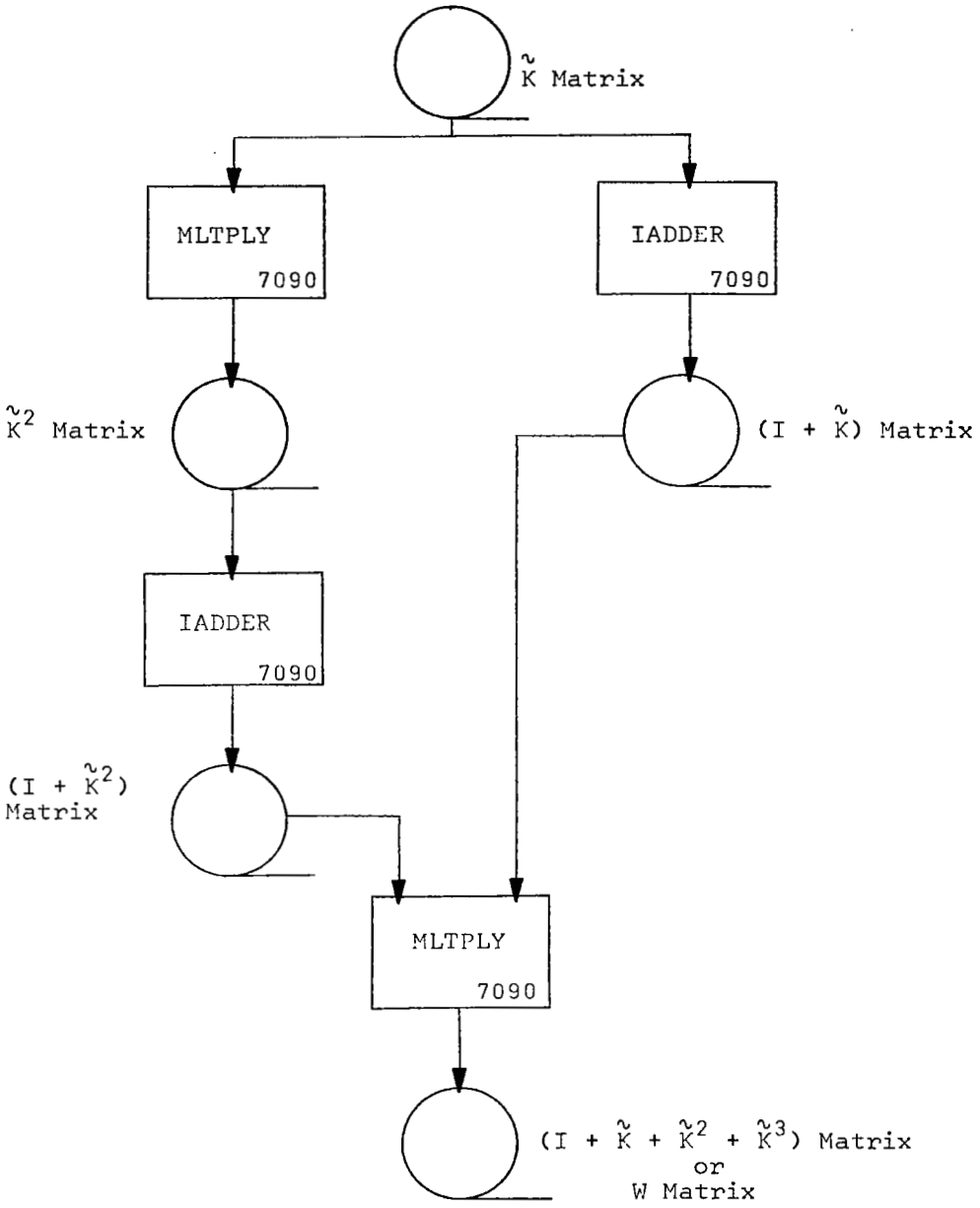


FIGURE 29 ASSOCIATIVE MATRIX GENERATION (PART 2 of 3)

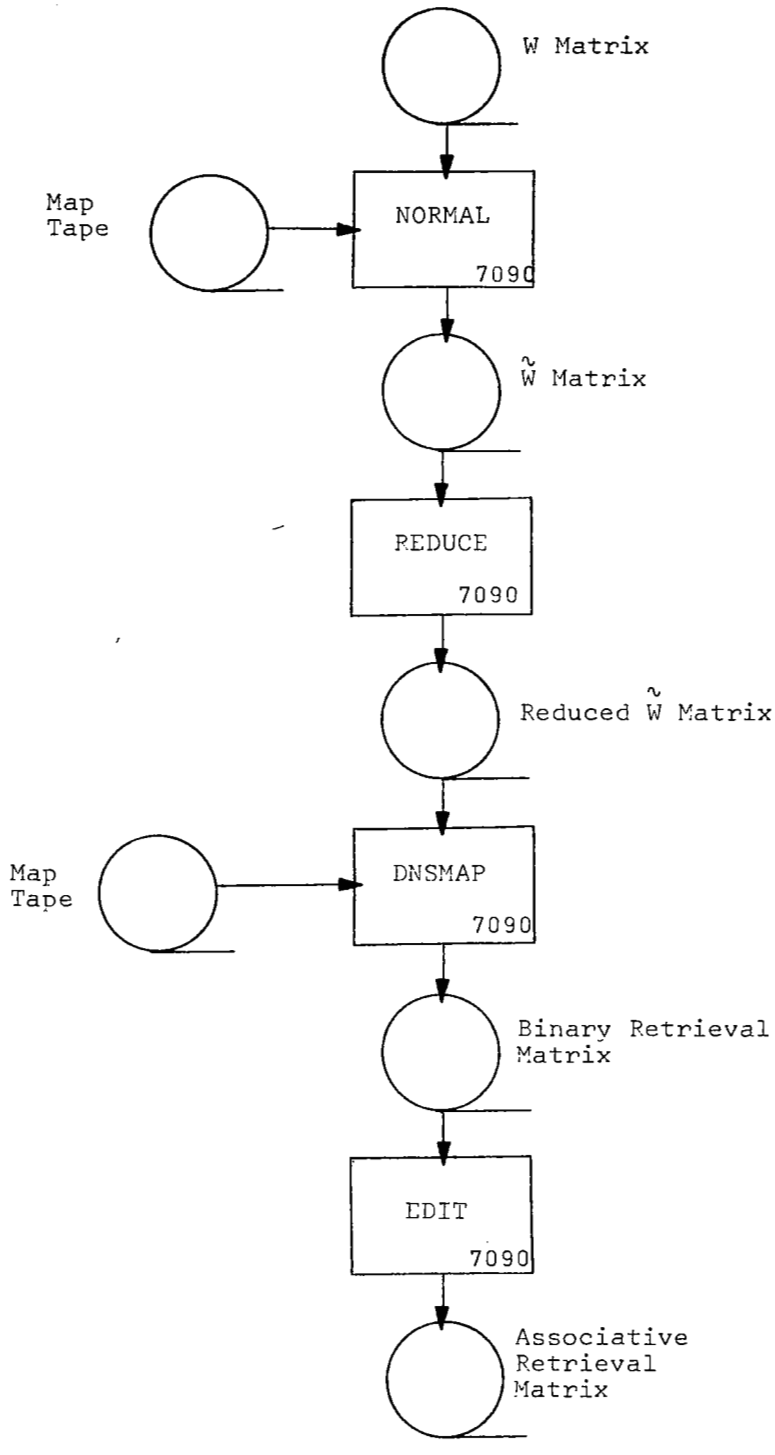


FIGURE 30 ASSOCIATIVE MATRIX GENERATION (PART 3 of 3)

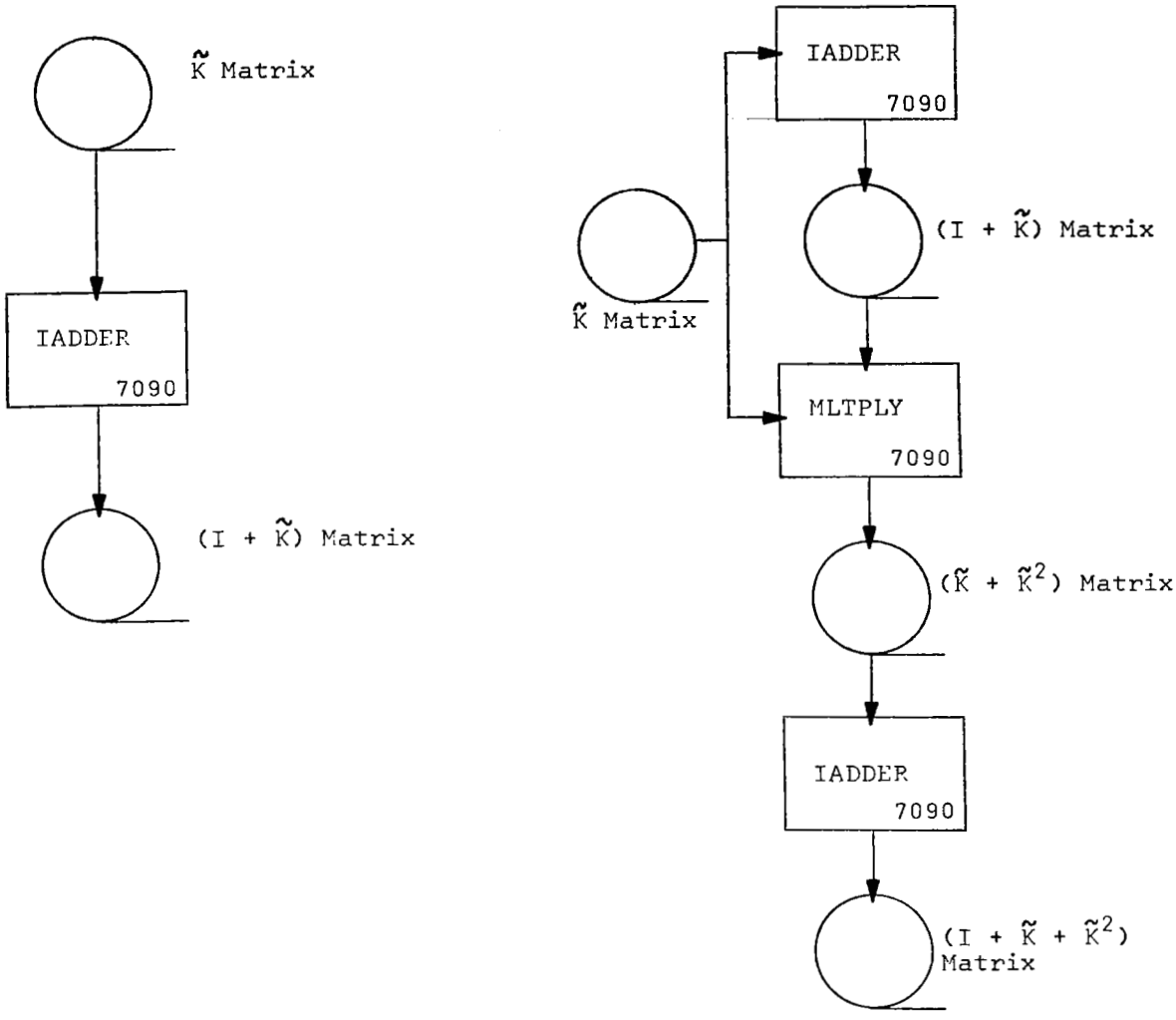


FIGURE 31 $W = I + \tilde{K}$ and $W = I + \tilde{K} + \tilde{K}^2$ Calculations

Once an approximation W to $(I-\tilde{K})^{-1}$ has been obtained, it is again normalized with NORMAL to keep common terms from exerting too great an influence.

Since the 1401 is a relatively slow computer, it is not practical tp pass a 1000x1000 matrix for a retrieval. In addition, most of the values are insignificant when compared to the higher values in a run. The matrix is therefore reduced in size with REDUCE by selecting the n most significant entries in each row of the matrix. The result is then passed through DNSMAP a second time where the original term numbers are re-inserted. Finally, the format is changed in EDIT so that this data can be accepted by the 1401.

B. Other Programs

Several other programs not previously discussed have been written for the NASADL system. These programs permit variations of the basic operations to be performed, and they also carry out general utility functions that have proven useful.

The associative matrix that is generated by the NASADL system always consists of the associations among the most frequently occurring terms and is, therefore, a square matrix. To facilitate the study of the profile of one or more segments of the matrix containing all the term co-occurrences, three programs, MODDNS, MODNRM and MODRDU, have been written. These are similar to DNSMAP, NORMAL and REDUCE with but small variations.

MODDNS permits the user to specify from one to nine frequency ranges. The program will retrieve the complete set of co-occurrence combinations for all terms which have occurred in the specified frequency regions. Thus, there can be formed a rectangular matrix of specified dimension in the row coordinate, with the number of terms in the NASADL file as the size of the column coordinate.

To carry out the functions of normalization and reduction, a different map is constructed since the row and column numbers cannot be condensed. These programs also allow for rows much longer than the main-line programs permit.

There is currently no mechanism for performing multiplication with the rectangular matrices. The addition of the unit matrix is still carried out with IADDER. However, this program will only add the unit value to the rows already present. The output from MODRDU is processed in the normal manner by EDIT.

A program, FRQPAR, has been developed to produce a file of row number/column number/matrix value entries from any matrix-formatted file. A sort to order these entries by column number is also available.

Lastly, two utility programs have been provided. PRINTR provides an easy-to-read listing of the contents of any matrix-formatted file or of the term-term frequency file. A seven-column print format conserves space. Decimal equivalents of the matrix values are printed. CPTRPA makes an extra copy of the term-term frequency file. This file is expensive to create. This program provides a backup file in the event that the original is accidentally destroyed.

C. 1401 Retrieval Program

The retrieval portion of the NASADL system can be carried out on either an IBM 1401 or an IBM 1410 computer. The retrieval run is made up of two phases. Phase 1, shown in Figure 32, accomplishes word association retrieval to extend the scope of a query. Phase 2, shown in Figure 33 performs the library search functions which result in document selection.

A user formulates a query by preparing a set of cards to be punched with term names and weights. He also provides a threshold to determine relevancy and some identification to be punched in a header card.

The retrieval program uses as input four files which are created and maintained by other phases of the system.

a. A dictionary file of all keywords (terms) in the library and the code number assigned to each.
b. A matrix file containing term-term associations, that is, a numeric representation of the strength of the relationship between all pairs of terms.
c. A matrix file containing document-term associations, that is, for each document the terms which the document has been indexed with.
d. A document abstract file which contains a brief description of each document in the library.

Communication from Phase 1 to the user consists of the following reports:

a. Request listing, providing a record of the input request.
b. Notification of unacceptable terms if any. Terms not in the dictionary are rejected.
c. "One-word association" listing and cards. A "one-word association" is a term profile, that is, a list of all terms associated with an input request term. In matrix terminology, all columns of the row which represents an input term are printed out as "one-word associations". These terms are also punched out at the same time. The profiles are produced only if the user requests them.
d. Term-association listing and cards. This report lists all terms found to have a relevance factor greater than the threshold. In addition to being listed, term associations are also punched on cards and stored on magnetic tapes.

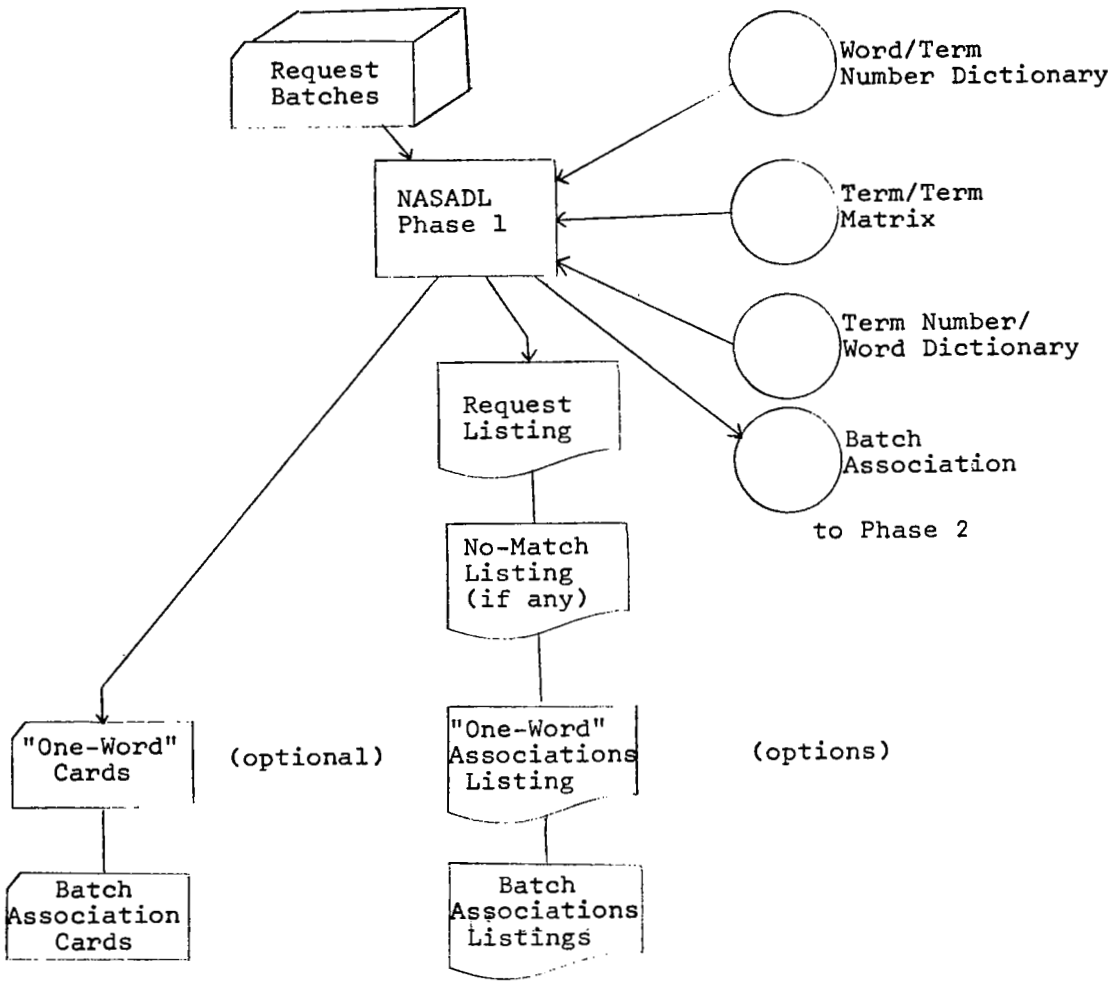


FIGURE 32 SYSTEM FLOW - PHASE 1 ASSOCIATIVE RETRIEVAL

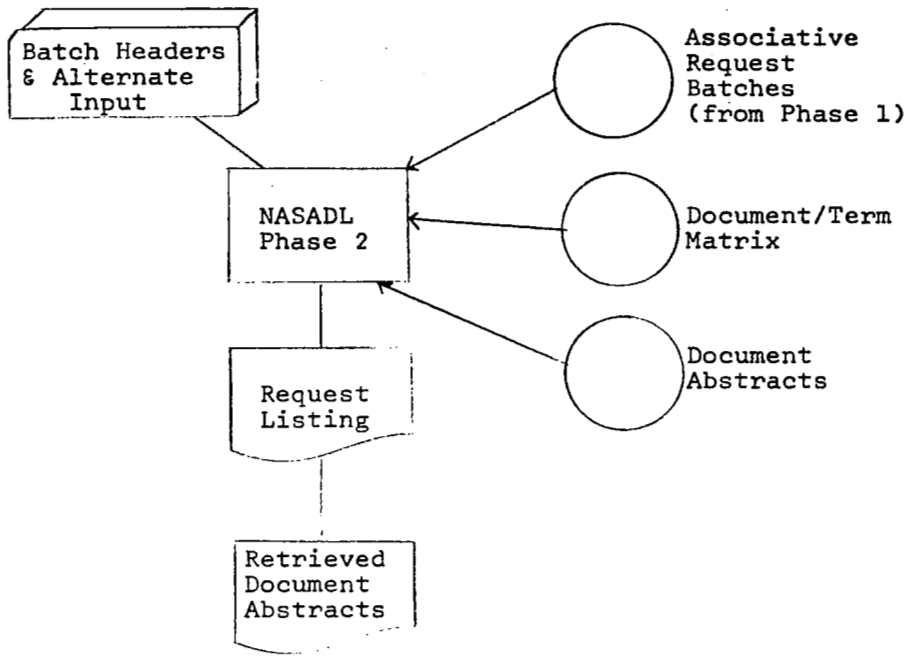


FIGURE 33 SYSTEM FLOW - PHASE 2 DOCUMENT RETRIEVAL

Phase 2 accepts input request terms either from cards or from a term-association tape produced by Phase 1. Unless specific action is taken to introduce card input, Phase 2 will normally proceed using tape input. Again a request listing is produced first. No other reports are printed by Phase 2 until the final report of retrieved document abstracts is issued. The abstract listing is ranked by relevance.

The system is set up to batch process several requests at one time. The maximum number of request sets is 50. There is no individual limit on the number of terms that can be included in a request set, however, the total number of terms in all sets is limited by the machine being used. For a 1401 with 8,000 characters of memory, the limit is 300. For a 16,000 character 1401 or a 1410, the limit is raised to 900.

1. Control Cards - The user formulates his initial inquiry by describing the subject in terms recognizable to the system. The terms recognized by the system are those included in the dictionary file. While the system vocabulary may be extensive, request terms should be checked against a listing of the dictionary file to prevent rejections because of differences in spelling and so forth. Each term must be weighted. Weights may range from +.9999 to -.9999. A term with a blank or zero weight will be ignored.

The user must supply some identification and a Phase 1 threshhold by means of a request header card. Accuracy to six decimal places is allowed in the threshold. The threshold is always positive. If one-word profiles are desired, the user must punch in a 1 in column 10 of the header. To aid the users, a request form such as that shown in Figure 34 might be employed.

Unless a user specifies otherwise, his request will be processed through both phases, and he will receive as output both term associations and document retrieval listings. Phase 1 will prepare a header card for Phase 2 with standard values specified as follows:

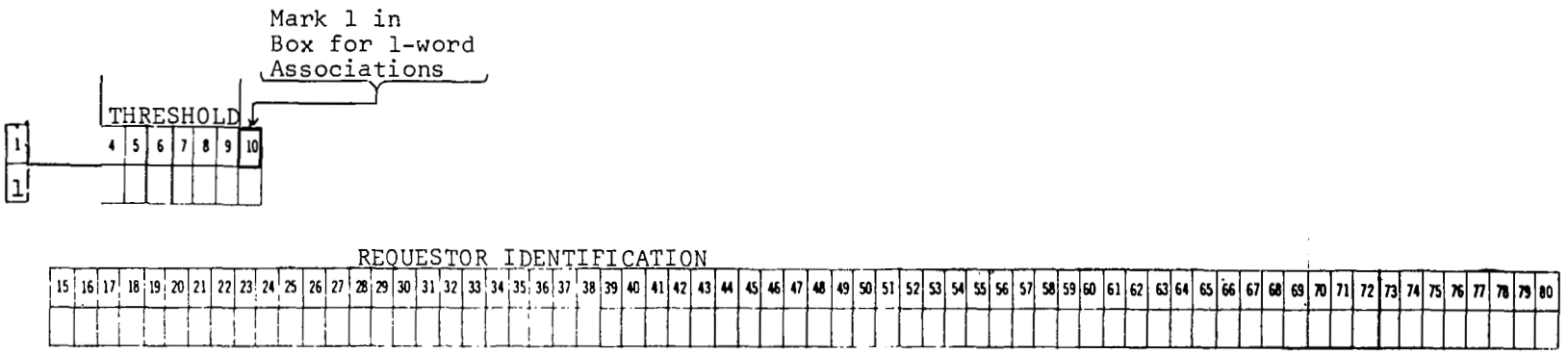
batch - each batch number assigned
input - will be tape (rather than card)
threshold - will be .001
input - limited to the top 50.

This card directs Phase 2 to process the request.

The user may, of course, submit a Phase 2 header card containing other than the above values. In addition to the listings the user will also receive from Phase 1 his request card deck, associative terms card deck, and a partially completed header card for Phase 2 (to be used if the user desires to run Phase 2 again). He may then formulate a document retrieval request using the cards produced from Phase 1. This deck may be directly entered into Phase 2 processing, provided that the header card is completed. The Phase 2 header must specify either card or tape input by C or T in column 2. If tape is used for input, the

Card 1

RETRIEVAL REQUEST FORM



Card 2

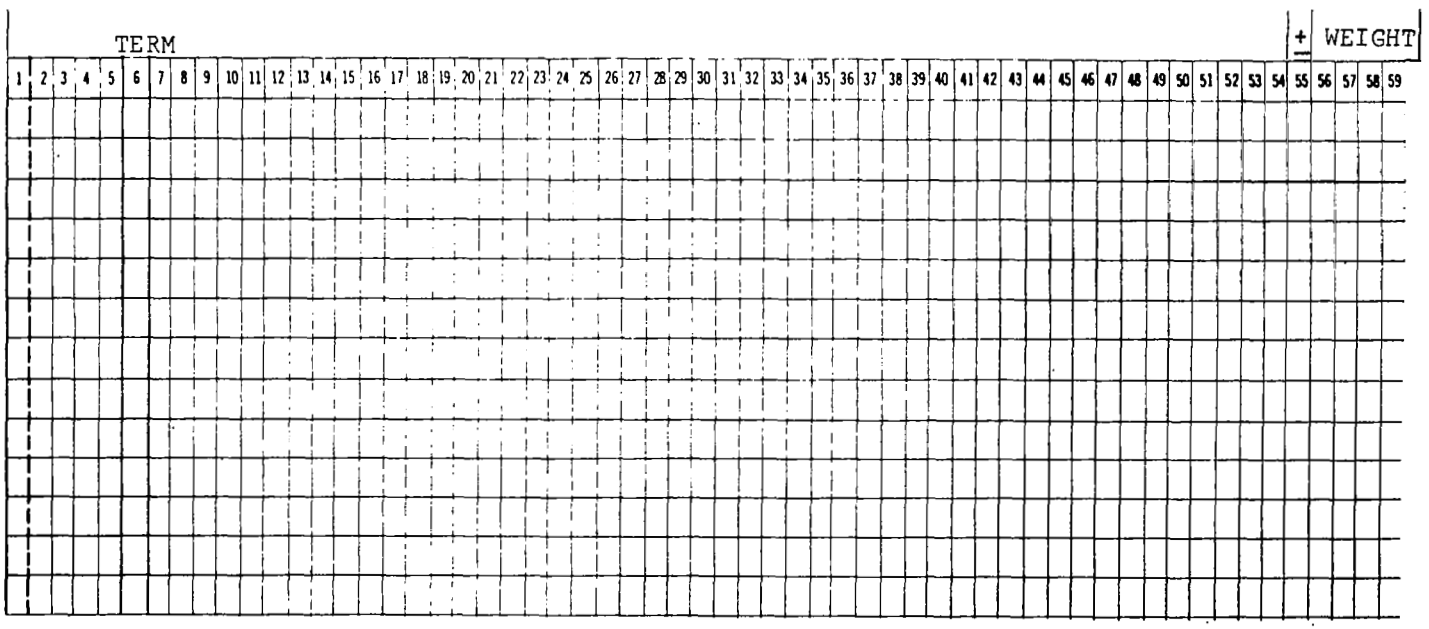


FIGURE 34 RETRIEVAL REQUEST FORM

tape to be used and the assigned batch number must be identified. Tape identification is the run identification assigned when the tape was written, and will be found on Phase 1 listings. The partially completed header card punched by Phase 1 contains tape and batch information, but does not carry user identification, threshold, or limits.

The Phase 2 header card may also specify a limit on the input, that is, if only the 20 highest ranked terms from Phase 1 associative retrieval are to be used for document retrieval, the user will supply an input limit of 20. If no limit is supplied, as many terms will be used as the machine's core storage capacity permits.

2. System Operation - Although a phase is made up of many programs, the operation of a phase, once it is started, is continuous except when tapes must be changed. When necessary, the program will print a message specifying what action is required. The number of halts for tape handling depends on the number of tape units in use. The system expects either 4 or 6 drives. With 6 tapes, Phase 1 is uninterrupted. Drive numbers 1 through 6 are addressed regardless of the number of drives available. For 6 drives, mount tapes as follows for Phase 1:

1. NASADLDICTAPALPHAB
2. NASADLMATRIXEDITED
3. NASADLDICTAPTERMNO
4. scratch
5. scratch
6. scratch

If 4 drives are available, mount tapes 1, 4, 5, and 6. When the alpha-sorted dictionary on drive 1 is no longer needed, a message is printed:

REPLACE NASADLDICTAPALPHAB TAPE ON 1
WITH NASADLMATRIXEDITED ON 2.

Similarly, when the matrix tape is completed, a message is printed directing the operator to replace it with the NASADLDICTERMNO tape on drive 3.

Initial assignments for Phase 2 are:

1. NASADLBCDDTFMATRIX
2. NASADLABSTRACTS001
3. scratch
4. scratch
5. Input from Phase 1
6. scratch

If only 4 drives may be used, mount tapes 3 through 6 and the program will print instructions as needed.

Several tapes produced by different runs of Phase 1 may be called. To avoid excessive tape handling, the operator should examine request header cards and group them by run identification, which is punched in columns 69 through 77.

In both phases, a run number must be assigned. A card is punched as follows:

Columns 1-6 - PHASE1 or PHASE2
7-14 - Run identification

This card is placed in front of data decks. The run identification is printed on all reports, punched in all output cards, and also is used as a label for the Phase 1 output tape which serves as input to Phase 2.

Request cards for both phases must be inserted in the program decks where marked. The card containing the run identification is placed in front of the first request. Each request set must have its header card first.

At the end of Phase 1, card disposition is as follows:

a. The rightmost stacker contains the program deck.
b. Input cards in the next stacker are to be returned to users.
c. The center stacker contains batch association cards with a partially completed Phase 2 header card in front of each batch. These cards should be interpreted and distributed.
d. The next stacker contains Phase 2 header cards for all batches with standard values. These cards become input to Phase 2.
e. The leftmost stacker contains blank cards.

To aid in separating batches of output in the center pocket, the operator should place blank cards behind the program deck in the card reader. The cards in the punch feed should be a contrasting color. The program will merge one blank card from the read side in front of each batch punched into the center pocket.

At the end of Phase 1, the output tape should be labelled and file protected. A message to this effect is printed.

NASADL Retrieval Program - Phase 1 Operating Instructions

Card Deck	-	Insert control card and input data into Phase 1 program deck where marked. (Place blank cards behind program deck. Use a different color than is in the punch feed.)
Punch	-	Insert blank cards and turn on.
Switches	-	Set Switch A and I/O switch ON. All other switches OFF.
Printer	-	Any paper, any carriage tape.
Tapes	-	Tape assignments will be printed out by the program.
Halts	-	1111 Tape error. Press START to try 10 more times.
		2222 Operator action required. Printer message specifies what action to take.
		7777 End of phase.

NASADL Retrieval Program - Phase 2 Operating Instructions

Card Deck - Insert control card and input data into Phase 2 program deck where marked.

Switches - Set switch A and I/O switch ON. All other switches ON.

Printer - Any paper, any carriage tape.

Tapes - Tape assignments will be printed out by the program.

Halts - 1111 Tape error. Press START to try 10 more times.
2222 Operator action required. Printer message specifies what action to take.
7777 End of phase.

BIBLIOGRAPHY

1. Stevens, M. E., Heilprin, L. and Giuliano, V.E., (Eds.), "Statistical Association Methods for Mechanized Documentation," NBS Misc. Pub. 269, Washington, D. C. (1964).

2. Arthur D. Little, Inc., "Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems," Report ESD-TR-66-405 (1966).

3. Stiles, H. E., "The Association Factor in Information Retrieval," J.A.C.M., Vol. 8, pp. 271-279 (1961).

4. Stiles, H. E., "Progress in the Use of the Association Factor in Information Retrieval," unpublished (1962).

5. Maron, M. E., Kuhns, J. L., and Ray, L., "On Relevance, Probabilistic Indexing and Information Retrieval," J.A.C.M., Vol. 7, pp. 216-244 (1960).

6. Doyle, L. B., "Semantic Road Maps for Literature Searchers," J.A.C.M., Vol. 8, pp. 553-578 (1961).

7. Arthur D. Little, Inc., Studies for the Design of an English Command and Control Language System, Report CACL-1 (ESD-TR-62-45) (1962).

8. Arthur D. Little, Inc., Studies for the Design of an English Command and Control Language System, Report CACL-3 (ESD-TR-63-673)(1963).

9. Arthur D. Little, Inc., Centralization and Documentation, Final Report to the National Science Foundation, C-64469 (1963).

10. Jones, P. E., "Research on a Linear Network Model and Analog Device for Associative Retrieval," in Automation and Scientific Communication, American Documentation Institute (1963).

11. Giuliano, V. E., and Jones, P. E., "Linear Associative Information Retrieval," in Vistas in Information Handling (Howerton and Weeks, Eds.), Spartan Press (1963).

12. Arthur D. Little, Inc., Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems, Report ESD-TR-66-405 (1966).

13. Fossum, E. G., et al., Optimization and Standardization of Information Retrieval Language and Systems, Final Report, Contract AF 49(638)-1194, UNIVAC, Blue Bell, Pa., (1966).

14. Sherry, M. E., "Memory Organization of a 7090 to do Statistical Association Processing for Document Retrieval," in Parameters of Information Science, American Documentation Institute, October 1964.

15. Giuliano, V. E., "Analog Networks for Word Association," IEEE Transactions on Military Electronics, MIL-7, No. 2, Sc. 3 (1963).

16. Doyle, L. B., "Indexing and Abstracting by Association," American Documentation, Vol. 13, (4), 1962.

17. Stevens, M. E., Automatic Indexing: A State-of-the-Art Report, NBS Monograph 91 (1965).

18. Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," IBM J. Research and Development, Vol. 1, (1957).

19. Cleverdon, C. W., Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems, ASLIB, Cranfield, England (1962).

20. O'Connor, J., "Automatic Subject Recognition in Scientific Papers: An Empirical Study," J.A.C.M., Vol. 12 (4), pp. 490-515, (1965).

21. Spiegel, J., Bennett, E., et al., "Statistical Association Procedures for Message Content Analysis," Information System Language Studies No. 1, Report SR-79, Mitre Corp., (1962).

22. Kuhns, J. L., "The Continuum of Coefficients of Association" in Statistical Association Methods for Mechanized Documentation, NBS, Misc. Pub. 269, (1965).

23. Dennis, S. F., "The Construction of a Thesaurus Automatically from a Sample of Text," in Statistical Association Methods for Mechanized Documentation, NBS, Misc. Pub. 269, (1965).

24. Dale, A. G., and Dale, N., Some Clumping Experiments for Information Retrieval, Report LRC-64-WP1A, Linguistics Research Center, U. of Texas, (1964).

25. Curtice, R. M., and Rosenberg, U., Optimizing Retrieval Results with Man-Machine Interaction, Center for the Information Sciences, Lehigh Univ., (1965).

26. Salton, G., "Some Experiments in the Generation of Word and Document Associations," in Proceedings of the Fall Joint Computer Conference, (1962).

27. Salton, G., "Associative Document Retrieval Techniques Using Bibliographic Information," J.A.C.M., Vol. 10, (4), p. 440, (1963).

28. Salton, G., "The Evaluation of Automatic Retrieval Procedures -- Selected Test Results Using the SMART System," American Documentation, Vol. 16 (3), 1965.

29. Salton, G., "An Evaluation Program for Associative Indexing," in Statistical Association Methods for Mechanized Documentation, NBS Misc. Pub. 269 (1965).

30. Salton, G., "A Combined Program of Statistical and Linguistic Procedures for Automatic Information Classification and Selection," in Automation and Scientific Communication, American Documentation Institute, (1963).

APPENDIX A

PRODUCING PROFILES WITH SECOND ORDER ASSOCIATIONS USING THE 1401 RETRIEVAL PROGRAM

ABSTRACT: The Linear Association Model expresses the desired term association matrix operator as a matrix series

$$(D + DKD + DKDKD + \cdots)$$

where D is a diagonal normalization matrix and K is a submatrix of the co-occurrence count matrix. The third term in the above series reflects the contribution of second order associations. This appendix documents the procedure to follow in using the 1401 retrieval program (Phase I) twice to discern the contribution of the second order associations.

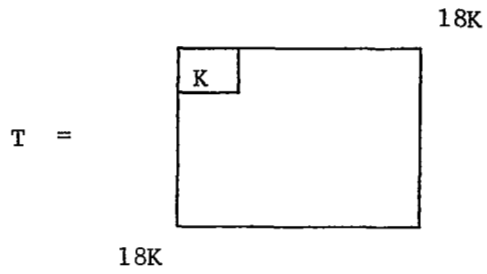
The procedure described is the most attractive method available at the present time for investigating second order associations for a small number of terms. Detailed examination of the effect of the second order associations for a few terms is immensely cheaper than squaring the whole matrix and then looking at the profiles.

A. Definitions

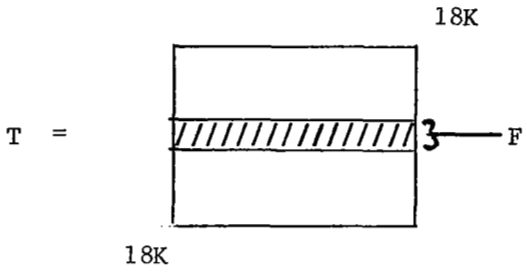
Let T be the symmetric matrix of co-occurrence counts, an 18K x 18K matrix in the NASA case. T_{ij} is a record of the number of documents containing both term i and term j where here, by convention, we regard

the terms to be numbered in decreasing order of usage frequency.

Let K be the 1K x 1K submatrix of co-occurrences among the 1000 most frequent terms.



Let F be any horizontal slice matrix, i.e., submatrix consisting of a number of rows of T.



Let D be an 18K x 18K diagonal matrix for which $D_{ii} = \frac{1}{f_i}$, i.e., a matrix which displays the reciprocal of the term frequencies along the diagonal.

B. The Retrieval Program

The operation of the retrieval program can be summarized most easily if we think of it as operating with a (virtual) 18,000 x 18,000 term association matrix A which can be made to be anything we please. Then if the query is represented by the 18,000 dimensional column vector q,

the program produces (in Phase I) the output vector $w = q'$, where

$q' = \frac{1}{(q^T q)} \cdot q$, i. e., a re-scaled version of q.

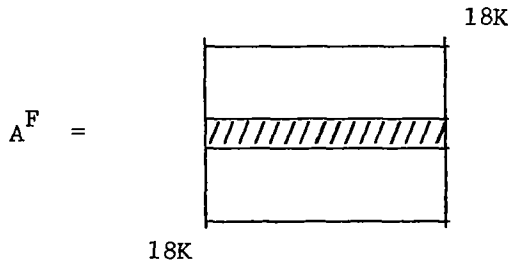
In addition to w, the program can provide on request, the rows of A for all terms with nonzero weight in q (which we call "one-word associations"). Let p^1 , p^2 , etc., denote these one-word association profiles.

C. Available Matrices

The association matrices A available for use by the retrieval program to date serve well to illustrate the reasoning which follows.

The first matrix shall be called A^F because it consists mainly of the information from the rectangular slice matrix F previously defined.

The matrix A^F looks like



where the shaded slice consists of the same rows we included in F. But rather than containing co-occurrence counts as F did, the rows are now normalized. Explicitly

a) if i is a shaded row, $A^F_{ij} = 100 \times \frac{f_{ij}}{f_i \cdot f_j} \geqslant 2^{-7}$

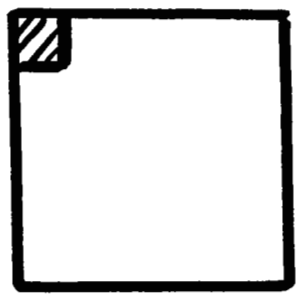
b) if i is not a shaded row, $A^F_{ij} = 0$

Thus, importantly, A^F is a submatrix of 100 • (DTD reduced), i.e., of 100 • DTD with all elements $< 2^{-7}$ thrown away.

Operating the retrieval program using A^F produces a "one-word association" profile p', for any term j whose row falls in the shaded region. We know that this profile is a row of 100 • (DTD reduced). But because this last matrix is symmetric, we can also regard p' as being the j^{th} <u>column</u>. Thus, (since reducing the matrices by throwing out the tiny elements is not a significant difference for our purposes) we may consider that p' above is the column vector $w^j = 100\ DTD\ q^j$ where q^j is the unit vector with 1 in position j.

Let us now consider the other matrix we have available, viz. the matrix A^K defined analogously:

$A^K =$



where

a) if i, j lie in the shaded region ($i, j \leq 1081$), then for $i \neq j$,

$$A^K_{ij} = \frac{f_{ij}}{f_i\ f_j}$$

provided $\frac{f_{ij}}{f_i\ f_j} > t_i$ (a variable threshold from row to row that leaves no more than 120 nonzero elements per row of A^K).

b) for $i = j$ in the shaded region $A^K_{ii} = \frac{1}{f_i}$

c) otherwise $A^K_{ij} = 0$

We see immediately that A^K contains an approximation to the upper left hand $1K \times 1K$ submatrix of $D + DTD$.

D. Summary

For preciseness, let us define two "selection matrices," $18K \times 18K$ diagonal matrices with ones and zeros on the diagonal.

$\$^F_{ii}$ = 1 if i is a term whose row is in the rectangular slice matrix, F, 0 otherwise.

$\$^K_{ii}$ = 1 if i is one of the top 1081 high frequency terms, 0 otherwise.

Then:

$$A^F = 100\ \$^F\ DTD$$

$$A^K = \$^K\ (DTD + D)\ \$^K$$

Normal operation of retrieval program with inquiry q produces

$$W = \frac{1}{q^T q} \cdot A_q$$

One word associations output of the retrieval program for a single term query q^j

$$W = A^T q^j$$

E. Producing the Profile with Second-Order Effects for Term j

1. Insert term j in the program and ask for one-word associations. Choose j so that A^F applies, i.e., $\$^F q^j \neq 0$

$$W = \left(A^F\right)^T q^j = \left(100 \cdot \$^F \text{ DTD}\right)^T q^j$$

$$= 100 \cdot \text{DTD}\$^F q^j$$

2. Multiply by $\$^K D^{-1}$

(Small FORTRAN program that multiplies output weights on terms by their usage frequency.)

3. Pose the resulting profile to the retrieval program again, using the standard request mode, and the matrix A^K.

Result: $$W' = \left(\$^K \text{ (D+DTD) } \$^K\right)\left(\$^K D^{-1}\right) \cdot 100 \left(\text{DTD}\$^F q^j\right)$$

$$= 100 \ \$^K \text{ (D+DTD) } \$^K \text{TD}\$^F q^j$$

But $$\$^F q^j = q^j$$

$$W' = 100 \cdot \$^K \text{ (D+DTD) } \$^K \text{TD} q^j$$

So far we have assumed, for the broadest generality, that $\$^K q^j$ can be = 0, i.e., that the query term of interest could fail to be one of the top 1000. (Our concern here, of course, is with getting a profile over the top 1000 that embodies the effect of the square of the 1000 x 1000 matrix.) To show how we obtain second-generation effects, we analyze the two cases -- first when the query term does fall in the K matrix, and next the case when it does not.

Case I:

Suppose, however, that

$$w' = 100\$^K (D+DTD)\ \$^K T\$^K D q^j$$

$$= 100\ \sqrt{D}\,(\sqrt{D}\ \$^K T\$^K\ \sqrt{D}\ + (\ \sqrt{D}\ \$^K T\$^K \sqrt{D}\ )(\sqrt{D}\ \$^K T\$^K \sqrt{D}\ ))$$

$$\cdot\ \sqrt{D}\ q^j$$

Let $\sqrt{D}\ \$^K T\$^K \sqrt{D} = \tilde{K}$

$$w' = 100 \sqrt{D}\ (\tilde{K}+\tilde{K}^2)\ \sqrt{D}\ q^j$$

And we note that the only difference between this and

$$100\ \sqrt{D}\ (I+\tilde{K}+K^2)\ \sqrt{D}\ q^j$$

is in the j^{th} component of the output, i. e., the self-association of j is not correctly calculated in w'. But the query term is usually artificially placed at the top of the list, so this is of little concern. Or we could manually calculate the difference, viz.

$$100Dq^j = \frac{100}{f_j}\ q^j$$

and add this into the j^{th} component of w' to obtain the exact weight term j ought to receive. Thus, the operation of the program in the stated way does include accurately the effect of squaring the association matrix.

Case II:

Let $\$^K q^j = 0$

Then it turns out that what we are getting is expressible in terms of a matrix which has one more row and column than $\tilde{K}$ in Case I. This matrix

is the result of pretending that we had increased the size of the submatrix K by one in order to include term j's co-occurrences with other high-frequency terms in the matrix.

Let $\$^j$ be a diagonal matrix with a 1 in position jj only. Then $\$^j q^j = q^j$ and $(\$^K+\$^j)q^j = \$^{K'}q^j = q^j$

Consider the expression

$$W'' = 100\$^{K'}(D+DTD)\ \$^{K'}T\$^{K'}Dq^j$$

As in Case I, this expression is convertible directly into the form

$$W'' = 100\sqrt{D}\ (I+\tilde{K}'+\tilde{K}'^2)\sqrt{D}\ q^j - 100Dq^j$$

which is, by inspection, the result of an operation on q^j that incorporates the effect of $\tilde{K}'^2$, with only the weight on term j in error. Note that now the matrix, K' , is the matrix we would have obtained if term j had been "wired into the network," i. e., chosen as an association term.

We consider ourselves satisfied if we can produce W'' using the 1401 retrieval program. But we have shown in Case I that the retrieval program produces

$$W' = 100\$^K\ (D+DTD)\ \$^K TDq^j$$

Using the facts

$$\begin{cases} \$^K = \$^{K'} - \$^j \\ \$^{K'}q^j = q^j \end{cases}$$

let us substitute in this expression to partition out various effects. Thus, rewriting:

$$W' = 100\ (\$^{K'}-\$^{j})\ (D+DTD)(\$^{K'}-\$^{j})\ T\$^{K'}Dq^{j}$$

$$= 100\$^{K'}\ (D+DTD)\ \$^{K'}T\$^{K'}Dq^{j}$$

$$- 100\$^{j}\ (D+DTD)\ \$^{K'}T\$^{K'}Dq^{j}$$

$$- 100\$^{K'}\ (D+DTD)\ \$^{j}\ T\$^{K'}Dq^{j}$$

$$+ 100\$^{j}\ (D+DTD)\ \$^{j}\ T\$^{K'}Dq^{j}$$

The first term is W".

The last term and the second (because of pre-multiplication of a vector by $\$^{j}$) consist of vectors with a single nonzero element in position j.

Thus, only the third term remains to be considered. Its contribution can be rewritten

$$-100\$^{K'}\ (D+DTD)\ \$^{j}T\$^{j}Dq^{j}$$

But $\$^{j}T\$^{j}$ is the selection of a diagonal element from T whose diagonal elements are 0. Thus, the third term contributes 0.

Accordingly, except for the $j^{\underline{th}}$ component,

$$W' = W"$$

Thus, we see that the use of the 1401 retrieval program twice -- first with a one-word profile operation using a hollow rectangular slice matrix A^F , then with the square matrix A^K -- provides all the information we need for discerning the effect of second order associations. The procedure is extraordinarily simple to accomplish.

"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Scientific and technical information generated under a NASA contract or grant and considered an important contribution to existing knowledge.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

TECHNOLOGY UTILIZATION PUBLICATIONS: Information on technology used by NASA that may be of particular interest in commercial and other non-aerospace applications. Publications include Tech Briefs, Technology Utilization Reports and Notes, and Technology Surveys.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION DIVISION

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Washington, D.C. 20546